Chapter 1

Detection: Overview of Historical, Societal, and Technical Issues

Lloyd A. Currie

Center for Analytical Chemistry, National Bureau of Standards, Gaithersburg, MD 20899

Practical societal needs and basic scientific advances frequently rely on Measurement Processes possessing specified detection capabilities with acceptable probabilities of false positives and false negatives. The first part of this overview introduces the basic concept of (chemical) detection, together with its applicability to selected societal problems such as the detection of natural hazards and the implementation of certain regulations. Basic scientific measurement issues concerning assumptions and their validity, plus hypothesis testing and decision theory as related to analyte detection are next introduced. Part two comprises a brief historical review, highlighting major contributions to the concept and realization of detection in chemical applications. The current state of the art is then considered. Part three is the most extensive, as it seeks to expose most of the technical issues involved in deriving meaningful detection decisions and detection limits, considering the overall Chemical Measurement Process. Those concerned primarily with societal or historical matters may wish to pass over this part. Among the topics discussed are: systematic and model error; non-normal random error; the special problem of the blank; replication vs Poisson variance; issues concerning complex data evaluation, calibration, and reporting -- including pitfalls associated with "black box" algorithms; OC curves; power of the t-test; and quality. The section concludes with some new material on discrimination limits, lower and upper regulatory limits, multiple detection decisions, and univariate and multivariate identification. A brief summary follows, bringing together historical, societal, and technical highlights. A concluding observation is that a meaningful approach to practical societal needs is at hand, but that order must be brought out of the extant diversity of technical views on detection.

The DETECTION LIMIT (L_D) is one of the most important characteristics of any Measurement Process. Recognizing the existence of such limits is crucial both for strictly scientific endeavors, such as the search for a new fundamental particle (1), and for vital societal applications of scientific measurements, such as the detection of a pathological state or a hazardous level of a heavy metal. In this latter regard, important progress has been made in conveying to the public and their policy makers that it is a **law of measurement science** that the detection capability of all Measurement Processes must stop short of zero, in close analogy with the Third Law of Thermodynamics.

Recognition that L_D may not be zero, has alleviated earlier legislative problems, such as the dictum that no residue of proven animal carcinogens may be present in certain food products (2). The fact, however, that detection limits can, at a cost and with technological advances, be made ever smaller has forced reexamination of regulatory issues in the light of extant and even potential detection capabilities. The consequence has been the consideration of cost/benefit or "acceptable risk" alternatives to "no detectable residue" regulatory policy (3). Such alternatives are mandatory in light of the fundamental principles of detection. Defining acceptable levels of risk (4), whether in a regulatory setting or with respect to medical decisions or even in terms of governmental actions in connection with potential natural disasters, is primarily a sociopolitical matter. Although this issue is of central importance, it transcends the theme of this chapter, which is to examine the historical evolution and current state of the art of detection from the perspective of chemical measurement science.

In order to highlight the importance of Detection Decisions and Detection Limits, and to underline the fact that the probability of detection does not immediately pass from zero to unity at the Detection Limit, we have presented in Fig. 1 several situations where valid detection decisions and adequate detection limits are of considerable practical importance. (The presence of a finite risk of error (false negative) at the detection limit -- i.e., the absence of "certainty" -- is the second aspect of the problem that is somewhat foreign to the common understanding, the first being the fact that zero detection limits are unattainable.) This figure introduces the Hypothesis Testing foundation for Detection, and it demonstrates that it is essential for those of us involved in measurement science to develop a sound, common, and quantitative approach to the formulation of Detection Limits. In addition, this formulation must be communicated in an effective manner both within the scientific community and with those who depend on our measurements for societal decisions and policy making.

As a final introductory note, it should be observed that from the perspectives of basic discoveries in Science and the early discernment of fundamental changes in the Global Environment (e.g.,

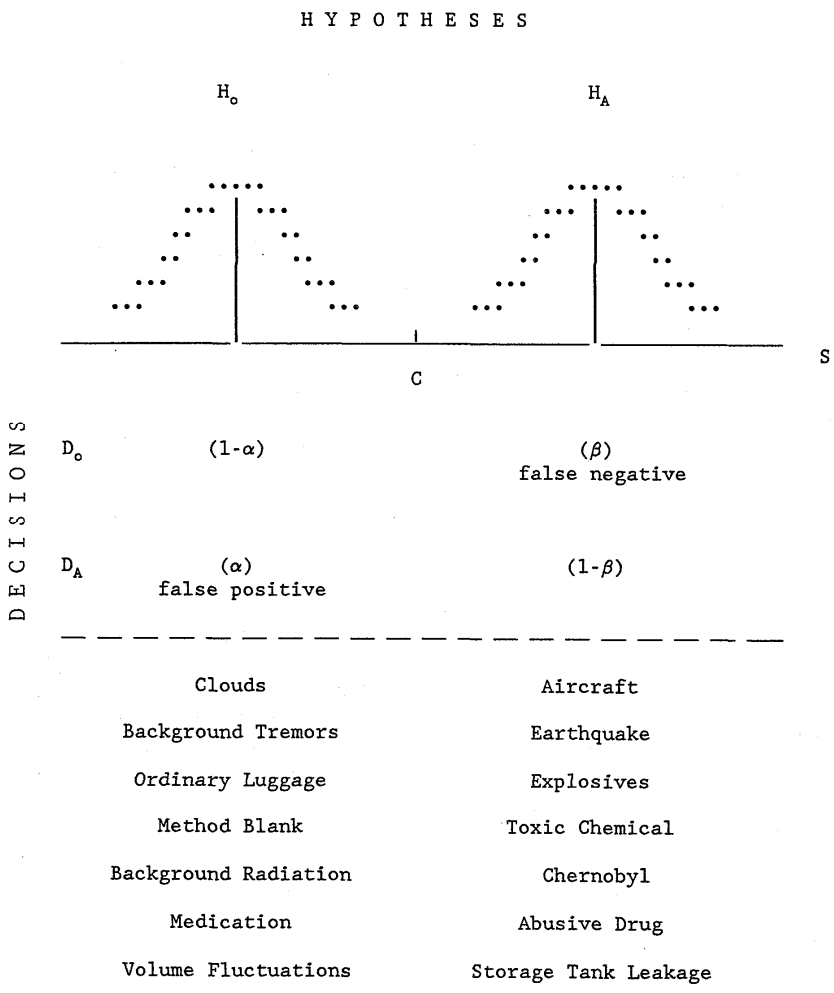

Fig. 1. Hypothesis Testing and Detection Limits. The upper part of the figure indicates the null [H_o] and alternative [H_A] hypotheses, with the corresponding decisions [D_o, D_A] at the left. Two kinds of erroneous decisions may be made: false positives [probability α] and false negatives [probability β]. (S represents a signal level; C, a decision point or "critical" level.) The lower section contrasts a number of "real world" H_o's and H_A's where adequate detection limits for the H_A's have clear, practical consequences.

stratospheric ozone changes, CO_2-induced global warming), the ability to **design** measurement processes having sufficient detection capability places one at the "cutting edge." Repeatedly in Science, one finds that discoveries are made just as the signals begin to emerge from the noise; and it is the "trained eye" which is generally the first to grasp them. Also, in the context of experimental design, it should be noted that *absolute* detection limits are often the goal, in that the hypothesis (or phenomenon) to be detected is generally conceived of in absolute rather than relative units.

1. BREADTH: The Scope of Detection

Fires, earthquakes and other natural hazards, pathological states, chemical contaminants, new fundamental particles or theories, instigators or sources of pollution or crime, natural or anthropogenic events of the past -- these are all illustrations wherein the basic concept of **Detection**, especially as embodied in the **Statistical Theory of Hypothesis Testing**, occupies a central position. Hypothesis formation -- i.e., specification of the source or system state or phenomenon to be tested [the "null hypothesis"] -- is necessarily the first step. For example, one might wish to test the null hypothesis (H_o) that no earthquake occurred (at a given time and place). To test H_o, one requires a test- or measurement-process (MP), often a Chemical Measurement Process (CMP), the outcome of which yields a **Decision** regarding the validity of the null hypothesis. The "alternative hypothesis" (H_A) which we wish to be able to detect -- e.g., an earthquake of a given magnitude -- must exceed the **Detection Limit** of the Measurement Process employed.

The keys to understanding the meaning of Detection Decisions and Detection Limits in matters of practical importance to science and society are: a) the existence of the two states [or hypotheses] which we wish to distinguish; b) a specified measurement process having an adequate DETECTION LIMIT; and c) a threshold or CRITICAL LEVEL for the measurement variable [Signal] for making the Detection Decision. Unfortunately, no measurement process can be exact, so false positives [α-error, e.g., earthquake erroneously "detected"] and false negatives [β-error, e.g., actual earthquake missed] will occur. Perhaps a more common example is that of the fire alarm. The measurement in this case might be made with a smoke detector, which if set to too low a threshold might give a false alarm [α-error] due to cooking fumes; if the critical level or threshold is set too high, a real fire of some consequence might be missed [β-error]. If an adequate balance between these two types of error cannot be achieved, one needs a **better measurement process** -- i.e., a detector having a lower detection limit. Note that the detection limit is an **inherent property of the measurement process**, whereas the detection decision is made by comparing an outcome or result of measurement with the Critical Level [threshold setting].

Fig. 1 suggests a wide range of situations where adequate detection limits are crucial for the well-being of society. The figure implies that the alternative hypothesis has a unique value on the x-axis. This is sometimes true. For example, the

radiocarbon concentration [isotope ratio] for living matter is $^{14}C/^{12}C = 1.18 \times 10^{-12}$, whilst that for fossil fuel carbon is effectively zero (5). A similar situation obtains for population means for chemical concentrations indicative of certain pathological states (e.g., glucose in diabetes (6)), or trace element concentrations characteristic of certain ore bodies. In a great majority of cases, however, the intensity or magnitude variable (x-axis) can take on many discrete (denumerable H_A's), or even continuous values (infinite number of H_A's). Such is the case, for example, with chemical or radioactivity contamination, earthquakes, fires, hurricanes, etc. For a given measurement process a special relation exists among the "distance" between H_A and H_o, and the two kinds of error, α and β. Fixing any two of these quantities determines the third, as will be shown in a later discussion of "Operating Characteristics." (Section 3.2.3.)

1.1 Regulatory Limits and Detection Limits. The practical significance of detection limits is best appreciated in connection with a specific external problem. Thus, based on quantitative assessment of health effects or of a new scientific phenomenon, one may conclude that it is vital to be able to detect a signal or concentration level as low as, say L_R. It follows that a measurement process having L_D no greater than L_R must be selected or, costs permitting, designed to meet the need. This is illustrated in Fig. 2 which depicts the critical level and detection limit schematically for earthquakes. The upper part of the figure presents an hypothetical relation between damage or societal cost and undetected earthquake magnitude, together with a maximum acceptable cost which fixes a "regulatory limit," L_R. (L_R might be defined, for example, by the "balance point" at which the false positive [false alarm] cost -- the cost of evacuation, is equivalent to the false negative cost -- damage incurred or lives lost in the absence of evacuation.) The lower part of the figure indicates the signal detection limit of a measurement process which meets this need. Also shown is the dependence of L_D and the two types of hypothesis testing errors on random measurement error. (The lower portion of the figure, for actual earthquake forecasting, relates to precursor measurement processes. The wealth of physical and chemical precursors utilized are reviewed by K. Mogi in Science, 1986, 233, 324.)

Two observations, perhaps obvious, follow from Fig. 2: first, a zero magnitude earthquake could not in principle be made detectable; second, with improving performance [decreased detection limit] formerly undetectable tremors will be found. Lack of appreciation of these fundamental principles of measurement may lead to regulatory difficulties, such as the requirement that any non-zero quantities of chemical carcinogens should be detectable, or that any detectable amounts should be reported (2). The latter has in effect been equivalent to a moving target, as analytical procedures continue to advance dramatically [Note 1].

A footnote on the matter of regulation, which leads directly to our next topic, relates to the relatively recent cost/benefit basis for regulatory decisions (3) and the emergence of the discipline of Risk Assessment (4). That is, that despite the lack of any explicit incorporation of a dollar value on human life in

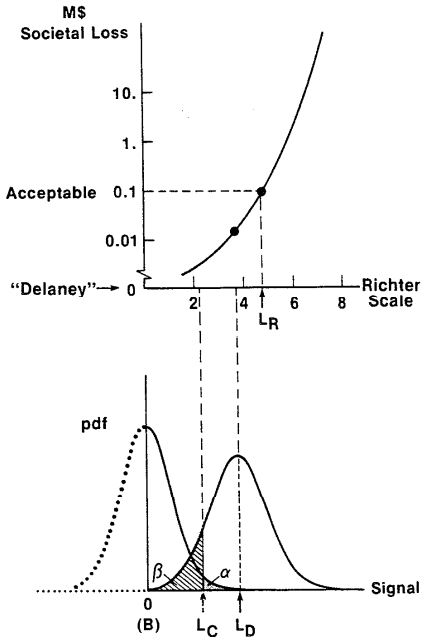

Fig. 2. Regulatory Levels [L_R] and Detection Limits [L_D]. The upper portion of the figure traces a presumed relation between earthquake magnitude [abscissa] and cost to society [ordinate]. The "Delaney amendment" viewpoint (not defined for earthquakes) might be interpreted as requiring zero societal risk and a corresponding L_R magnitude of zero, which of course is scientifically unattainable. Rather, an acceptable cost to society for undetected earthquakes, here imagined to be 0.1 M$, is used to establish the requisite "regulatory" level. The lower part of the figure represents the corresponding earthquake measurement process or precursor alarm (seismograph signal, radon emanations, biological [animal] sensors, etc.). The requisite DETECTION LIMIT [L_D] must now be no greater than L_R, and L_D in turn is related to the probability density functions [pdf] for the null signal [H_o: S=0] and the signal to be detected [H_A: S=L_D], and acceptable false decision probabilities α, β. L_C is fixed by the H_o-pdf and α; L_D is then set by L_C and β, given the H_A-pdf.

the algebra of regulation, a de facto "$2 million unwritten rule" has evolved (7). An analysis of 10 years of regulatory decisions in the US relating to chemical carcinogens showed that this value fairly consistently marked the point above which regulations were classified as too costly to impose, and below which regulations were judged as warranted.

Our use of the symbol L_R, incidentally, is not restricted to regulatory matters. Earthquakes, for example, cannot be regulated! Rather, L_R denotes the external limit which drives the design of our measurement process. It could apply as well to the requirements of a high quality production process, or a tracer study of long range atmospheric transport, or the investigation of extremely slow reaction processes, or in fact any of the situations indicated in Fig. 1. Perhaps it might better be labeled "reference limit (or level)" or "requisite limit."

1.2 Decision Theory and Societal Decisions/Actions. The foregoing introduction to detection theory was based strictly on the Neyman-Pearson or "frequentist" approach to significance testing and signal detection (8,9), with the exception of the imposition of an external reference or regulatory limit, L_R, based on sociopolitical and/or scientific considerations. An alternative approach, especially appropriate for (detection) decisions culminating in some kind of action, is provided by the application of Decision Theory (10), or more generally Decision Analysis (11). Although this theory may be of considerable importance for certain societal or business decision-making, its structure is such that it is not generally applied to chemical measurements.

The major advantages of the decision theoretic approach are that it permits one to apply explicit loss functions to the erroneous decisions [α,β-errors], and that it readily incorporates prior (or "subjective") knowledge concerning the probabilities of the respective hypotheses. The ability to utilize loss functions and prior probability is advantageous in that costs and beliefs and values external to the measurement process may be effectively incorporated into the decision making. A complication is that there may not be unanimity concerning the weights to be assigned to these quantities; this is somewhat analogous to the complications in reaching agreement on appropriate values for L_R. [Costs, for example, would doubtless be viewed differently by regulators and regulatees, producers and consumers, physicians and patients, etc. The issue is analogous to the question of "whose experts" are speaking in Court or advising in Congress -- i.e., it is necessarily tempered by advocacy positions.] Except when one is treating a strictly scientific question, however, it is important to realize that the losses and prior probabilities are frequently complex sociopolitical and/or economic matters, best determined by experts in those fields.

Decision theory operates on the basis of an "objective function" which is in some way optimized through the setting of a decision threshold. A lucid presentation to alternative strategies for formulating detection decisions has been given by Liteanu and Rica (8, p. 192). The essence of the matter is that a threshold value k_o for the Likelihood Ratio is derived from a) prior probabilities for the null and alternative hypotheses, b) a cost or

loss matrix specifying costs associated with correct and erroneous decisions, and the probability density functions (pdf) for experimental outcomes for each of the hypotheses in question. These data are combined to compute the mean loss (or cost or risk) which is then minimized in order to derive k_o. The decision test is performed by comparing the observed (experimental) value for k with k_o. [k, the likelihood ratio, is the ratio of the pdf for H_A to that for H_o at the signal level in question.] The optimal value, k_o based on the "Bayes Criterion" is given by the product of the net cost of a false positive and the prior probability of H_o divided by the product of the net cost of a false negative and the prior probability of H_A. An interesting illustration leading to the same conclusion is given in Massart, Dijkstra and Kaufman (12, p. 516) in connection with medical diagnoses and selection of the optimal point on the Receiver Operating Characteristic Curve [ROC]. [Operating Characteristic (OC) and ROC curves will be discussed briefly in a subsequent section.] The issue of developing an "optimal" decision strategy based on the prior distributions of both well and ill patients is an interesting one. In the illustration presented in Ref. 12 (pp. 508 ff), for example, there is a presumed preponderance of healthy patients [prior distributions]. By using the distributional crossing point as the threshold, one finds that about half of the abnormal (ill) subpopulation would have been misdiagnosed! (See also Appendix H in Egan (9) for an interesting illustration of the Bayesian approach to medical decision making, and the consequent need for multiple diagnostic tests -- a non-trivial issue in the light of current efforts of major medical insurers to curtail the number of diagnostic tests.)

References (8) and (10) give alternative decision strategies -- Minimax, Ideal Observer, and Maximal Likelihood -- when only partial information is available for prior probabilities and/or costs. [The Minimax approach, for example, cuts one's losses from a wrong guess for the prior probabilities.] The effects of special preferences or aversions [e.g., to extreme cost] are discussed in terms of "Utility Theory" by Howard (11), as well as the use of Decision Analysis for designing sequential experiments and the setting of research priorities.

This brief excursion into Decision Theory is included to indicate the manner in which experimental data can be coupled with external (societal) judgments to form a logical basis for societal decisions and actions. A justification for so complex a strategy for decision making is that "simple" scientific measurements and model evaluations will always be characterized by measurement uncertainty. Yet societal decisions and actions must take place even under the shadow of uncertainty. For scientific measurements, as discussed in the following text, however, we shall restrict our attention to the relatively simple Neyman-Pearson hypothesis testing model (8, p. 198).

1.3 Testing of Assumptions. The detection of erroneous assumptions lies at the core of sound measurement science. It is therefore especially appropriate to include reference to Detection Decisions and Detection Limits for key assumptions in our survey of the scope of Detection. Assumptions of principal importance for chemical measurements include those relating to the functional form

and parameters for a) the physicochemical (or empirical) model and b) the error model relating the experimental observations to the underlying chemical composition. Among the assumptions, or assumed parameters to be tested, the following are of special importance:

Functional Relation
- o number of chemical components
- o characteristic spectra or chromatographic patterns
- o mathematical relation for the response for each component (includes correct identification, and curve shape)
- o matrix effects and interference [interactions] among components
- o parameters such as the blank, recovery, sensitivity (efficiency)

Error Model
- o cumulative distribution function [type]
- o parameters [variance, higher moments] (variance components for compound distributions)
- o autocorrelation [non-white noise]
- o systematic error or bias [bounds]
- o blunders (discrimination from chance outliers, from discoveries)

Hypothesis testing is applicable to all of the above factors. Detection decisions may be made, for example, using the critical level of Student's-t to test for bias, or the critical level of χ^2 to test an assumed spectral shape or calibration model or error model. For a given measurement design and assumption test procedure, one can estimate the corresponding detection limit for the alternative hypothesis, e.g., the minimum detectable bias. As with analyte detection, the ability to detect erroneous assumptions rests heavily on the design of the experiment; and the study of optimal designs is a field unto itself.

A survey of several of the above model-parameter assumptions, as related to chemical component (or analyte) detection will be presented later. Let us terminate this preview with two observations: a) Tests of assumptions may themselves rest upon assumptions -- an obvious case being the use of Student's t, which rests upon the assumption of normality; b) Detection of an analyte through model failure (lack of fit) -- e.g., evaluating χ^2 when fitting a spectrum with one component missing -- is less sensitive than direct detection using the correct model. This is due to collinearity among spectral patterns (or overlapping chromatographic peaks) (13).

1.4 Analyte Detection. This is a primary focus for this volume, the specification of critical levels or thresholds for analyte detection decisions, and the design of CMP's to achieve requisite analyte detection limits. The following section includes an historical perspective on the topic. A tutorial is provided in the chapter by Kirchmer (14), where a crucial distinction is noted: that is, the detection decision is made in reference to an observed, random experimental outcome (estimated concentration),

whereas the detection limit refers to the underlying true concentration which the CMP is capable of detecting. The chief reason for interest in the latter is advanced planning and design -- i.e., assessing the capability of the CMP in question to meet the measurement needs.

Because of the broad scope of detection, as outlined in the preceding paragraphs, it is useful to distinguish some of the quantities or events detected with appropriate symbols. For the purposes of this chapter, the following will be used:

	Critical Level	Detection Limit
generic symbol	L_C	L_D
event or system state (earthquake, oil spill)	θ_C	θ_D
analyte concentration (or amount)	x_C	x_D
instrument response (net signal)	S_C	S_D
bias	Δ_C	Δ_D
external random error (non-Poisson; "between")	σ_{xC}	σ_{xD}
model - lack of fit	$\chi^2_c, \ldots$	--

In addition to the above, L_R is used to denote the external limit which drives the design of the Measurement Process (MP). Thus, if successful process control, or early warning (natural or human disasters), or fundamental chemical research depends on achieving a limit L_R, then the MP must be so designed that its $L_D \leq L_R$.

Note that the critical level of the appropriate test statistic ($z_{1-\alpha}$, $t_{1-\alpha}$, etc) can generally be used as a normalized alternative to x_C, S_C, etc. The "detection limit" for a test statistic, however, is meaningless, as x_D, S_D, etc. refer to the true underlying quantity. A corollary is that the term "detection limit" is also without meaning in the absence of an alternative hypothesis. (This is perhaps an obvious philosophical matter, but in principle, the null hypothesis cannot be rejected, except by chance [α-error], if no alternative exists; the β-error is then necessarily undefined. Of course an unexpected rejection can lead to an exciting search for the alternative.)

2. HISTORICAL PERSPECTIVE

The dual questions, "How little can I detect?", and "Has something been detected?" have long caught the attention of analytical scientists. Throughout recent history (i.e., 20th century) a number of responses have been formulated, such as

- o The intuitive [formulation]: basing detection decisions and limits on sound, but not readily quantifiable experience

- The ad hoc: selecting a rigid formula, often based on some reasonable limiting condition, via dictum, voting or consensus
- The signal/noise: generally assuming white noise, and addressing primarily testing of an observed signal
- The avoidance: only results or measurement processes thoroughly removed from the detection limit deserve our attention
- The hypothesis testing: where explicit attention is given to the risks of both false positive and false negative detection decisions.

In reviewing the history of detection limits (in Analytical Chemistry) it is helpful to keep these several, often implicit, differences in mind. If it is agreed that the concept of detection has meaning, then it is essential that the above questions be fully defined and explicitly addressed. In the view of this author a meaningful approach to analyte detection must be consistent with our approach to uncertainty components of measurement processes and experimental results; the soundest approach is probably the last [hypothesis testing] tempered with an appropriate measure of the first [scientific intuition].

Table I has been prepared from this perspective. The authors selected are drawn primarily from those who have contributed basic statements on the issue of detection capabilities of chemical measurement processes ["detection limits"], as opposed to simply addressing detection decisions for observed results ["critical levels"]. In fairness to those not listed, it is important to note that a) a selection only, spanning the last several decades has been given, and that b) there also exist many excellent articles (15,16) and books (12,17,18) which review the topic. It is immediately clear from Table I that the terminology has been wide ranging, even in those cases where the conceptual basis (hypothesis testing) has been identical. Nomenclature, unlike scientific facts and concepts, can be approached, however, through consensus. The International Union of Pure and Applied Chemistry [IUPAC], which appears twice in Table I, is the international body of chemists charged with this responsibility. At this point it will be helpful to examine the position of IUPAC as well as the contributions of some of the other authors cited in Table I.

Fritz Feigl (19), the father of "Spot Tests," heads the list primarily as one who suggested lower limits for chemical measurement, here translated (from the german) as "identification limits," which represented the best experience [or chemical intuition] of the day. Such limits, typically in the microgram range, were scarcely ad hoc, but they of course lacked the statistical sophistication of latter day limits. Feigl's limits, however, deserve our attention even today, in that they recognize the overall capability of the measurement process including that which cannot be readily treated by statistics [Note 2].

Table I. Historical Perspective -- Detection Limit Terminology

Feigl ('23)	-	Identification Limit (19)
Altshuler ('63)	-	Minimum Detectable True Activity (21)
Kaiser ('65-'68)	-	Limit of Guarantee for Purity (20)
St. John ('67)	-	Limiting Detectable Concentration (S/N_{rms}) (24)
Currie ('68)	-	Detection Limit (23)
Nicholson ('68)	-	Detectability (25)
IUPAC ('76)	-	Limit of Detection (29)
Ingle ('74)	-	("[too] complex...not common") (27)
Lochamy ('76)	-	Minimum Detectable Activity (92)
Grinzaid ('77)	-	Nonparametric Detection Limit (26)
Liteanu ('80)	-	Frequentometric Detection Limit (8)
NRC ('84)	-	Lower Limit of Detection [28]
IUPAC ('86)	-	Detection Limit (30)
IAEA ('87)	-	Detection Limit (93).

Among the others cited in Table I, Kaiser (20) deserves major credit for introducing the hypothesis testing concept into spectrochemical analysis, as does Altshuler (21) in radioactivity measurement. Wilson (22) championed its use for water analysis, and Currie (23) provided an approach for detection and quantification in analytical and radiochemistry. The reference by St. John (24) has been one of the most cited of those based on signal/noise, though it does not address the error of the second kind (false negative). Nicholson (25) gave one of the earliest treatments for extreme low-level (Poisson) counting data, and Grinzaid (26) offered a robust treatment not requiring the assumption of any specific distribution. Liteanu's frequentometric method (8) was also distribution-free, in the sense that an experimental estimate of the detection limit was derived from the **observed** fraction of false negatives, using a regression technique. The paper by Ingle (27), which was obviously designed to be tutorial (published in the J. Chemical Education) is noteworthy in that it suggested that the concept of the error of the second kind (which is intrinsic to the statistical theory of hypothesis testing) was simply too complex for ordinary chemists to grasp! Regrettably, there seems to be some support for such a statement; but Hypothesis Testing is one of the keystones of every elementary course in Statistics, so its formal introduction into the education of the analytical chemist would seem not too esoteric a step.

An exhaustive review of the definition and application of Detection Limits for nuclear and analytical chemical measurements was published in 1984 (28). The reader may wish to scan the titles of the papers there cited, to gain further insight regarding basic principles and terminology, counting statistics, non-counting and non-normal random errors, random and systematic variations in the blank, Bayesian approaches, reporting, averaging and censoring treatments, optimization, influence of alternative spectrum deconvolution techniques, etc. In the body of Ref. 28 special attention is given also to topics such as simple and multicomponent nuclear spectrum fitting and extreme low-level counting.

With respect to IUPAC, both the position published in 1976 (29), which addressed nomenclature, symbols and units in analytical

optical spectroscopy, and the more general analytical nomenclature document, now in review (30), treat detection from the hypothesis testing viewpoint. A non-conceptual difference lies in the choice of the risk level (false positives and negatives). The 1976 report, which grew out of Kaiser's work, used a fixed value of 3.00 for the standard deviation multiplier (S/N) for detection decisions. This would correspond to a false positive risk of 1-0.9986, or 0.14% (1-sided test), if the population were normal, and σ known. (The false negative risk β was not explicitly treated.) The current IUPAC Nomenclature Document recommends risk levels (α, β) of 5%, corresponding to a multiplier of 1.645 for σ known, normal population. Both documents recognize the effects of varying degrees of freedom in estimating the variance of the blank; the latter document specifically recommends the use of Student's-t to compensate, just as is done in the construction of normal confidence intervals.

The historical evolution of this topic has resulted in some very unfortunate and needless confusion in both terminology and concept. **Awareness of the nature of this confusion is crucial**, if we as analytical scientists are to arrive at a common and meaningful approach to detection, an approach that can serve society rather than add an extra level of confusion to a topic which the public regards as already complicated, albeit important.

The facts are that for at least the last decade or two there has been broad international support for the hypothesis testing framework for making analyte detection decisions, and evaluating -- especially for purposes of design and planning -- the inherent detection capabilities of measurement processes. In this context, a number of authors and institutions have employed terms like "detection limit" (or "limit of detection") to denote the latter, inherent detection capability, generally in units of concentration or amount (8,12,17, 21-23, 25,30,36). In the earliest work of some who most strongly came to support the hypothesis testing model, however, the notion of the false negative [β - error] did not appear (31). Kaiser in particular labeled his threshold level "Die Nachweisgrenze," or Detection Limit. In 1965 Kaiser treated the second kind of error (β), and introduced "Die Garantiegrenze für Reinheit" as the corresponding true concentration level for the alternative hypothesis (32). Kaiser's impact on the field of Analytical Chemistry has been extremely significant, and it is not surprising that many chemists have adopted his terminology for the Detection Limit. It has been adopted, however, in many cases to indicate not just the signal/noise level for making detection decisions, but also as a measure of the inherent detection capability of the measurement process in question. Since the error of the second kind [β] exists whether it's recognized or not, this practice has led to a de facto false negative risk of 50% -- a value which is totally out of balance with a false positive risk of 0.14%, or even 5%! Thoughtful and lucid critique of this matter may be found in Ref's. 8 (p. 263) and 14. A curious footnote to this discussion is that one scarcely ever encounters Kaiser's second term, "Limit of Guarantee for Purity," in the scientific literature.

On the subject of nomenclature, a word concerning historically used terms for the detection *decision point* or level is in order. As stated immediately above, a number of analysts, following Kaiser, use "Limit of Detection" or "Detection Limit" as *both* the measure of (true concentration) detection capability and as a statistical critical level or threshold to make detection decisions. Following established practice in Statistics, the term "Critical Level" was recommended in (23). "Criterion of Detection" has been employed by Wilson (22); and Liteanu (8), who speaks of the "decision criterion" as a strategy, terms the numerical comparison level the "Decision (or Detection) Threshold."

The great majority of the authors cited in the foregoing discussion emphasized that the detection limit must refer to the entire analytical measurement process. In many cases one finds that not the case -- i.e., workers may refer (sometimes appropriately and intentionally) to just the instrumental measurement step, or to ideal, pure solution detection limits -- both of which may be far too optimistic for real, complex samples. Some compilations indicate "typical" detection limits, an acceptable practice provided the measurement process and sample nature (including matrix and interference effects) are rigidly controlled and subjected to appropriate ruggedness testing.

2.1 Present State of the Art. A perusal of the analytical literature two decades ago revealed considerable disparity in the specification of detection limits. This is shown in Fig. 3 which is reproduced from (23). Then current definitions spanned nearly 3 orders of magnitude when applied to the same measurement problem! Concern for such definitional (and/or conceptual) disparity has led a number of national and international organizations to address the need for a common, rational basis for treating this matter. Because of concentration related effects of trace chemical species on health, properties of high purity materials, and even global climate, relative detection limits for different measurement processes are not enough; detection capabilities must be assessed in absolute units. Awareness of the confusion surrounding detection limit practices, by organizations such as IUPAC, IAEA, ACS, a number of US regulatory agencies, and more recently CODATA (Committee on Data for Science and Technology) is a very positive thing. The difficulty and importance of the task is highlighted by several of the authors in this volume, notably: a) Crummett (33) ["In spite of extraordinary efforts (on the part of scientific societies to properly define detection limits) analysts continue to present their results in forms which cause the credibility of the data to be questioned or the meaning to be misinterpreted"]; b) Brossman (34) ["Attempts by our task force on low-level data to make a rigorous conceptual and statistical comparison ... have been unsuccessful. Even similar terms are defined in different, non-comparable ways..."]; and c) Currie and Parr (35), where it was observed that international interlaboratory comparisons involving the same bioenvironmental reference materials resulted in mutually exclusive results. For example, quantitative results for arsenic (in horse kidney) were reported by some laboratories at levels which exceeded the "detection limits" of other laboratories (which detected no arsenic) by as much as 4 orders of magnitude!

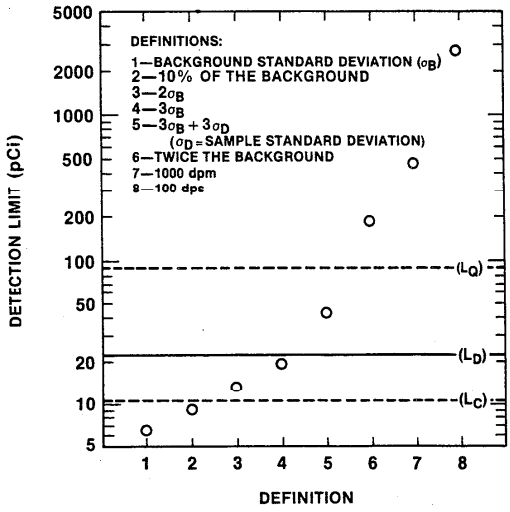

Fig. 3. Ordered detection limits -- 1968. (Reproduced from Ref. 23. Copyright 1968 American Chemical Society.)

Incidentally, the Brossman task force epitomizes a very serious problem: the coding of data into computerized data bases. Such data bases will doubtless have significant distributions and lifetimes, so the effects of possible distortions and information loss will be unfortunately amplified.

Understanding and acceptance of the hypothesis-testing position taken by IUPAC (29,30), the US Nuclear Regulatory Commission [28], the UK Water Research Centre (36), the IAEA, and reflected in many of the recent texts in Analytical Chemistry and Chemometrics (37), promises to resolve the needless, current disarray. Some of the current diversity can be seen in Fig. 4, which presents four of the principal detection limit definitions in vogue (and/or in regulatory guides) in the U.S. Comparisons among the statements, together with the supporting documents, show that: (1) the β error (false negative) is ignored in all but one, causing it to assume a de facto value 50%; (2) treatment of the blank is ambiguous or absent in two of the definitions, and restricted to the reagent blank in a third; and (3) a uniform approach to the α error, taking into account the number of degrees of freedom and Student's t is lacking. There is some irony in the fact that the fourth definition states that the LOD is "the lowest concentration ... statistically different from a blank", in view of a comment in the reference cited (Long and Winefordner, 1983). These authors note that the "well-based but seldom used concept in the calculation of detection limits ... the limit of guarantee for purity, c_G, described by Kaiser ... [represents] the lowest statistically discernable signal." Long and Winefordner go on to show that the original IUPAC definition (29) and the LOD of reference (16) indeed yield 50% false negatives (β). Kaiser's c_G, incidentally, is conceptually identical to Currie's L_D (23) and Boumans' "Limit of Identification" (15).

In addition to the above, one continually finds varying ad hoc or even undefined usage in the peer-reviewed analytical literature. For example, three recent papers, examined because of interest in their chemical content, all deemed detection limits of sufficient importance to include numerical tabulations. However, the first author stated that his detection limit represented a signal to noise ratio of 10; the second defined it as twice the standard deviation of the background signal; and the third gave no indication as to his meaning.

In conclusion, it is urgent that the analytical community adopt a uniform and defensible approach to the concept of detection. Apart from ad hoc or unstated procedures, failure to recognize the error of the second kind [β] -- i.e., failure to distinguish between detection decisions and detection capabilities -- is the most serious conceptual fault, placing false negatives at the level of coin-flipping accuracy. Failure to take into account all major sources of error, especially the nature of the blank, is the most serious measurement fault. A review of some of the more critical assumptions and technical issues related to valid detection limits follows.

Lower Limit of Detection (LLD). "The LLD is defined, for purposes of these specifications, as the smallest concentration of the radioactive material in a sample that will yield a net count, above system background, that will be detected with 95% probability with only 5% probability of falsely concluding that a blank observation represents a "real" signal" (94).

Instrumental Detection Limit (IDL). "The concentration equivalent to a signal, due to the analyte, which is equal to three times the standard deviation of a series of ten replicate measurements of a reagent blank signal at the same wavelength" (95).

Method Detection Limit (MDL). "The method detection limit (MDL) is defined as the minimum concentration of a substance that can be measured and reported with 99% confidence that the analyte concentration is greater than zero and is determined from analysis of a sample in a given matrix containing the analyte" (96).

Limit of Detection (LOD). "The limit of detection (LOD) is defined as the lowest concentration level that can be determined to be statistically different from a blank. The concept is reviewed in [ref. 38] together with the statistical basis for its evaluation. Additional concepts include method detection limit (MDL), which refers to the lowest concentration of analyte that a method can detect reliably in either a sample or blank, and the instrument detection limit (IDL), which refers to the smallest signal above background noise that an instrument can detect reliably. Sometimes, the IDL and LOD are operationally the same. In practice, an indication of whether an analyte is detected by an instrument is sometimes based on the extent of which the analyte signal exceeds peak-to-peak noise" (16).

Fig. 4. Four Definitions for Detection Limits Related to Current U.S. Regulatory Practice.

3. DEPTH: Limitations, Assumptions, and Technical Issues

The foregoing text represents a brief overview of some of the societal, historical, and broad conceptual issues relating to detection and chemical measurements. Here we offer an overview, in catalog or dictionary format, of a series of technical issues directly related to the estimation and validity of analyte detection limits. Balanced coverage has been the intent, but special attention has been given to topics not covered elsewhere in this volume, and to questions arising in discussions or put by "users" of detection limits. In some cases, this led to the introduction of new material such as multiple decisions and probabilistic pattern detection, utilization of physical constraints (on variance), and some effects of varying probability density functions [pdf] as related to experimental design and σ variation. The discussion is divided into three parts: the first, considering issues affecting the validity of detection decisions [null hypothesis testing]; the second, considering the type II error [β] and detection for the alternative hypothesis; the third, considering multiple detection decisions and Discrimination Limits for chemical species and chemical patterns. A guide to the topics presented in this section is given in Fig. 5 [Note 3].

3.1 Null Hypothesis Testing -- Assumptions and Conclusions. When an experiment is performed, we test the experimental result ($\hat{x}$) by comparison to the critical level or threshold (x_c) to decide whether or not analyte has been detected in excess of the blank or background level. Quite apart from questions involving the alternative hypothesis or detection limit, two crucial points must be kept in mind concerning the nature and validity of such a test. The first is that the **Statistical Test for Significance**, at significance level α, is based on **exactly** the same principles as the more fashionable calculation of **Confidence Intervals**, at confidence level 1-α (39). (In both cases, of course, one must pay attention to 1- vs 2-sided tests or intervals.) The second point follows: that assumptions affecting the validity of experimental confidence intervals are just as important for the validity of significance tests. Assumptions which demand our attention include the following: control [i.e., existence] of the measurement process; possible systematic model [functional] or measurement error; and properties of the random component (or components) of error -- i.e., form of the distribution [pdf or cdf], parameters of the distribution [σ^2,...], "color" of the noise (or noise power spectrum), and non-stationarity (changes with time) and heteroscedasticity (changes with concentration or other experimental parameters). Errors may arise also from uncompensated changes in the measurement process itself, such as alteration of the calibration [functional] model or random error model [cdf] due to chemical matrix effects or interference. No less important are the computational and data reporting strategies; these represent intrinsic parts of the **overall measurement process**. A brief catalog of selected hypothesis testing and experimental result related issues follows.

Detection Decisions, α-error [3.1]
(assumptions, validity)

- BASIC ERROR ISSUES
  - systematic error [3.1.1]
  - normal random error [3.1.2-.3]
  - non-normal [3.1.4-.5]
  - paired comparisons [3.1.6]
- MEASUREMENT PROCESS ISSUES
  - background, baseline, blank [3.1.7]
  - error components, truncation [3.1.8-.10]
  - evaluation process, calibration [3.1.11-.12]
  - reporting low-level data [3.1.13]
  - artificial thresholds [3.1.14]

Analyte Detection Limit, β-error [3.2]
(estimation, power)

- DETECTION LIMITS AND POWER
  - ignorance of β-error [3.2.1]
  - lower, upper L_D's [3.2.2]
  - α, β connection (ROC) [3.2.3]
  - power of the t-test [3.2.4]
- UNCERTAINTY IN L_D [3.2.5-.6]
- SPECIAL TOPICS
  - optimization [3.2.7]
  - multicomponent detection [3.2.8]
  - random error variation [3.2.9]
  - quality (algorithms, controls) [3.2.10-.11]

Discrimination Limit; Multiple Decisions [3.3]

- DISCRIMINATION LIMITS
  - lower and upper regulatory limits [3.3.1]
  - impurity detection [3.3.2]
- MULTIPLE DECISIONS, IDENTIFICATION
  - multiple detection decisions [3.3.3]
  - multichannel identification [3.3.4]
  - multivariable patterns [3.3.5-.6]

Fig. 5. Topical Guide to Technical Overview.

3.1.1 systematic and model error. Bounds for uncompensated, non-random errors must be allowed for by a corresponding increase in the critical level or confidence interval. This makes α an upper limit if the systematic bounds, which need not be symmetric, are given as upper limits. The validity of the corresponding uncertainty interval clearly depends heavily on the Chemical Intuition or scientific expertise employed, for example in identifying the range of possible alternative models.

3.1.2 normal (white) random noise. If σ is known, $L_C = z_{1-\alpha}\sigma_o$, where σ_o represents the standard deviation of the estimated net signal (23). If "simple" detection [gross signal - blank] is involved, where the blank is estimated from n equivalent observations, and the gross signal from one, then $\sigma_o = \sigma_B \sqrt{(n+1)/n}$ -- σ_B being the standard deviation of the blank. [Note that zero adjustment, as for the null level or baseline of a (recording) galvanometer, chromatograph, spectrophotometer, etc., does not eliminate the need for this blank or baseline estimate error propagation. Of course, σ_o may be scarcely greater than σ_B when a large span of linear baseline is quite precisely adjusted, for example, by least squares fitting or by graphic or even "eyeball" subtraction. A related point: the (detection) test must be applied to the net signal, or equivalent concentration estimate, because only that has an expected value under the null hypothesis of exactly zero. Imprecise knowledge of the mean value for the blank distribution prevents a rigorous test being applied to the gross signal. See comments below on the background, baseline, and blank.]

3.1.3 σ unknown (normal). Student's t replaces z when σ is estimated by replication. L_C now equals ts. [Note that L_C here, unlike L_D, is no longer a constant, because s (estimate of σ) is a random variable. Note also that σ (or s) as used in this text refers to the standard deviation of the "final" signal; if signal averaging or least squares fitting is employed to arrive at the final signal, then this should be interpreted as the standard error.]

3.1.4 non-white noise. The autocorrelation function (or spectral power density) must be taken into account in calculating critical levels or confidence intervals. This is not a trivial matter, and is remarkably often ignored. Its importance is seen most often for time dependent phenomena [e.g., in chromatography], and where "flicker noise" is found. Note that noise of this sort sets limits to the gains which may be achieved through signal integration or averaging. Note also that detection limits based on the Signal to Background Ratio derive from the assumption of background-carried flicker noise dominance. See especially Smit (40) and Epstein (41) for important discussions of this topic. In a broader sense the underlying issue relates to the limit in the information content of sets of observations which are not fully independent. One encounters it also when interpreting uncertainties for count rate meters (RC signal averages - (28, p. 96)), and uncertainties in functions of partially correlated random variables [error propagation, including covariance: (42)]. The topic is of special relevance when considering instrumental baselines [see below].

3.1.5 non-normal distributions. This problem can be dealt with rigorously if one knows the form of the random error distribution. A notable example of this occurs in "counting" experiments (e.g., radioactivity), where the physics of the process implies Poisson statistics. As the Poisson distribution is discrete, y_C (the critical level for gross counts) takes on integer values only, and α is generally in the form of an inequality -- ie, $\alpha \leq 0.05$ (28). Distribution-free techniques, especially those based on order statistics (such as the median and its confidence interval), and transformation techniques (eg, for log-normally distributed errors), are often appropriate (26,57). So-called non-parametric techniques -- the Gauss or Chebyshev inequalities, give (2-sided) α's as no greater than $(2/[3k])^2$ and $1/k^2$ respectively, where the standard deviation multiplier k replaces z of the normal distribution. Note that the Gauss Inequality is applicable for random variables having unimodal, continuous, and symmetric density functions, whereas the weaker Chebyshev Inequality is valid for any distribution having finite mean and variance. A small problem in applying the inequalities is that k must multiply σ, and σ is not generally known. Although s^2 is an unbiased estimate for σ^2 even for non-normal distributions, bounds for s/σ are distribution dependent and therefore also not generally known. [Note 4.] Difficulties are compounded when the measurement process consists of two or more steps comprising different kinds of pdf's. [See for example, Johnson, Ref. (44).]

Recommended solutions for the non-normality problem are: 1) use the percentage points of the actual pdf, if known; b) transform to normality; c) use order statistics; d) design the experiment to take advantage of "pairing" and the Central Limit Theorem. The last approach, which looks very attractive for chemical research, will be discussed below. Information on the other approaches may be obtained from specialized statistical texts (45).

3.1.6 paired comparisons; Central Limit Theorem. The Central Limit Theorem makes quality control charts work. Here, one charts sets of averages of observations and checks for excursions beyond Normal control limits. The averaging is done not primarily for standard error reduction, but to assure (approximate) normality. It can be shown that averages (or sums) derived from of a sequence of mutually independent random variables having a common distribution tend toward normality, often rather quickly (by the time n = 3 or 4). This "Central Limit Theorem" is valid regardless of the shape of the initial distribution, so long as it has finite variance. The rate of approach to normality, however, depends on the initial shape, being faster for symmetric distributions (45). For low-level chemical measurements, all too often the blank, which forms the basis for the detection decision, is neither symmetrically nor normally distributed -- especially when the blank is due to environmental or particulate contamination (46). Very wrong trace analytical confidence intervals and detection decisions may result. By designing the measurement process so that proper paired comparisons can be made (45, Chapt. 4), one can at the same time achieve the best statistical sensitivity and force symmetry (for the estimated net signal), and thus set the stage for approximate normality for averages of such estimates. Two other major reasons for forcing symmetry in this way are: a) to take

advantage of the median and its confidence interval for robust estimation, and b) to make possible the use of the Gauss Inequality for distribution free interval estimation. This approach has been suggested, for example, as a possible solution to the severe detection discrepancies obtained in IAEA intercomparisons (35). Specifically, the recommendation is to make detection decisions by comparing the averages of at least n=4 paired comparisons (gross signal - equivalent blank signal) with $ts/\sqrt{n}$, where s^2 is the estimated variance of the n net signals. (Note the direct analogy between the chemical blank measurement, and the "control" observation which plays a central role in clinical and psychological null hypothesis testing.) In anticipation of the next two topics, two related advantages of proper pairing (or equivalent fitting) may be stated: a) bias associated with systematic B-changes is minimized by taking "local" differences (y - B) and using local values for recovery and instrumental detection efficiency; b) effects of imprecision associated with certain "external" random variations (e.g., "between-day") may similarly be avoided.

3.1.7 background, baseline, blank. The variability of the null signal [B] is the determining factor in making valid detection decisions (H_o tests) or in deriving valid confidence intervals for low-level signals. In the ideal interference free, "pure solution" measurement environment, the instrumental background is the ultimate limiting factor. In this sense an "instrumental critical level" (decision level, threshold) and the corresponding detection limit mark the best possible performance of a system. Two cautions are in order, however. First, the instrumental noise (background variability) may not be white -- i.e., there may be long- or short-term variations which must be compensated for by appropriate modeling and/or astute paired comparisons. Second, when the chemical, physical, or geometric configuration of the final sample changes, there may be corresponding changes in the effective background (for example due to changes in external scattered radiation). Instrumental detection efficiencies or responses may also be perturbed by such sample-related factors (47), but that is a "calibration" matter, to be taken up separately.

Multicomponent instrumental responses, unless totally selective, generate spectral or chromatographic baselines which arise from complex physicochemical phenomena ranging from multiple (sample) particle or radiation scattering to component tailing, depending on the specific analytical technique involved. Such baselines generally subsume any instrumental background, and thus become the limiting factor. Valid net signal estimates and detection decisions then become critically dependent on accurate modeling of the baseline functional shape and noise structure. Empirical baseline shape models are common, the linear model being most used. Deviations from linearity may be modeled using low order polynomials or splines, but since the modeling is empirical one must be alert to possible model errors, such as unanticipated fine structure (48). The noise structure of the empirically modeled baseline deserves special attention in the case of drift, preferred periodicity, or more general autocorrelation such as "1/f noise" (40,41,49).

For all real chemical measurements, the chemical blank is the actual limiting factor. To assess its magnitude and variability,

there is no better approach than to apply the entire measurement process to an adequate number of real blanks. Unfortunately, this ideal may not be realizable, as it requires samples which are identical to those of interest in all respects except for the absence of the target analyte. The alternative approach is to attempt to "propagate" the components of the blank for each step of the CMP, taking into account the points of introduction, and subsequent recoveries and CMP-induced variations. This topic is enormously important and enormously complex. One must consider simultaneously: the effects of multiple blank sources; analyte, blank, and interferant recoveries for each CMP step; and instrumental detection efficiencies plus matrix effects for each (35, 50-52). A small complication arises from the fact that different types of blanks may exhibit different pdf's. Reagent and sample preparation blanks tend to be normally distributed, while environmental blanks are frequently log-normal (44,46,53). In the final analysis, of course, actual variations in the blank are convolved with instrumental noise. Lack of independence or normality for either will be reflected in the final, effective blank distribution. A final observation: systematic or random error in the estimated blank affects not only detection limits and confidence intervals for "low-level" samples; it may also limit the accuracy of high precision, "high-level" samples.

3.1.8 internal vs external error: propagation vs replication. The uncertainties of low level concentration estimates may be derived from error propagation for each stage or step of a compound CMP, or they may be deduced from replication and comparison with low-level SRMs for the overall measurement process, as in laboratory intercomparisons. Consistency between the two approaches is essential for the uncertainty estimates to be considered valid. Among the error characteristics that may be exposed through such ANOVA type testing are: "excess" random error [σ_x], systematic error [Δ], and covariance among internal errors. The first (σ_x) might represent, for example, a between day or between lab variance component, or in the case of Poisson counting statistics, an excess or non-counting component of random error. The second (Δ) could be manifest as the difference between the limiting mean for an intralaboratory measurement and the true (e.g., SRM) value. The third might be seen if internal, non-white noise or other error correlation effects were improperly accounted for in error propagation; the external estimate, derived from independent replicates would automatically compensate for such (internal) behavior. Comparison of internal and external estimates has its limitations, however. This is shown in Fig. 6 [$(\sigma_x,\Delta)_D$ vs n] which displays the detection limits for σ_x and Δ as a function of the number of replicates (54). Thus, in order to detect bias equal in magnitude to the standard deviation, one needs at least 12 degrees of freedom (13 replicates). To detect an extra variance component equal in magnitude to the known internal precision [$(\sigma_x)_D=\sigma_i$], one needs 46 degrees of freedom (55). These error components [σ_x, Δ] must, however, be taken into account if valid detection decisions are to be made.

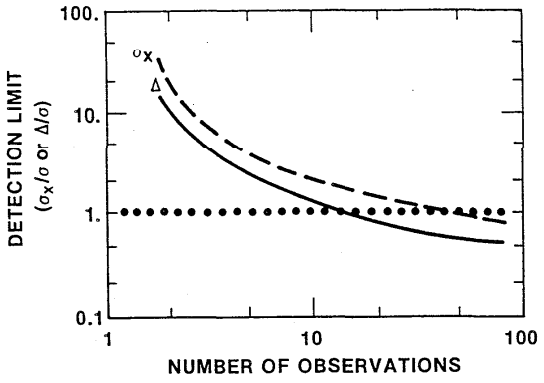

Fig. 6. Detection Limits for Bias [Δ] and excess (external) random error [σ_x] vs the Number of Observations. (Adapted from Ref. 54.)

3.1.9 internal error precisely known: improved detection decisions (and confidence intervals) using inequality constraints. If the excess random error component is well known, then obviously error propagation can be applied -- e.g., $V_t = V_i + V_x$ -- to calculate the total variance V_t, which is the quantity that must be used for calculating the critical level or confidence interval. (For notational simplicity, V is used here to denote variance, in place of σ^2. If multiplicative rather than additive relations are involved V would represent relative variance.) Quite frequently we find it relatively inexpensive to obtain precise estimates of V_i (internal variance) whereas external replicates -- e.g., between laboratories or between days, etc., -- tend to be more costly, hence fewer. Perhaps the extreme case of this sort occurs with Poisson counting statistics, where $V_i \approx N$ where N is the observed number of counts. If expected value of N is sufficiently large (e.g., >60) the expression yields a reasonably good estimate for V_i, as good as would result from 100 or more replicates. If V_x is taken to be zero, the critical level may then be calculated directly from V_i which here is estimated as the number of counts [Note 5].

If V_x is not known, we have three alternatives. One of extreme conservatism would be to use the lower and upper limits for V_x, based on replication [s^2=est(V_t)] and knowledge of V_i. Intermediate, and most common, is simply to calculate L_C as t•s, where t is based on the external number of df (number of replicates minus one). More interesting is the use of our knowledge that $V_x \geq 0$, and a variance weighted t. (A fourth alternative, ignoring the possible existence of V_x is all too common; the unsupported assumption that counting statistics or other internal instrumental variance fixes the overall imprecision can generate L_C's and CI's that are too small and correspondingly large false positive probabilities.)

The merit of the variance weighted t technique is that it permits us to use our excellent knowledge of V_i, together with the fact that V_x cannot be negative to obtain a significantly smaller L_C or CI than we would using s^2 directly. It provides protection against unanticipated external random error with little penalty if that error component is in fact negligible. The technique suggested here is tentative and approximate, but it appears to be conservative and asymptotically correct [Note 6].

To illustrate, let us consider triplicate measurements of a sample using a counting technique, such as ion counting mass spectrometry or photon counting in optical spectrometry. X-ray fluorescence analysis or gamma ray spectrometry. Internal variance derives from Poisson counting statistics [V_i] where the appropriate value of t_i equals its normal limit z_i or 1.645 for $\alpha = 0.05$. Total variance [V_t] for the 3 replicates is estimated as s^2, where t_t is 2.92 for 2 df. Excess variance [V_x] is $V_t - V_i$, and estimated as $s^2 - V_i$, with the constraint that V_x may not be negative. L_C' (or CI) is calculated as $t's'/\sqrt{n}$ where:

$$s' = \sqrt{V_i + V_x} \geq \sigma_i, \text{ i.e., } s' = \max(\sigma_i, s) \quad (1)$$

$$t' = t_i (V_i/V_t) + t_t (V_x/V_t) \quad (2)$$

For the example at hand, we estimate V_x as s^2-V_i. (Eq. 1) thus yields $s'=\sigma_i$ if $s \leq \sigma_i$, or $s'=s$ if $s > \sigma_i$. (Eq. 2) becomes $t'=1.645\ (k) + 2.92\ (1-k)$ where $k = \sigma_i^2/s^2$. For example, if σ_i is equivalent to 1.75 ng-Ca, and s, to 3.04 ng-Ca, k would equal $(1.75/3.04)^2 = 0.331$. As a result, $s' = 3.04$ ng-Ca, $t' = 2.50$, and $L_C' = 4.39$ ng-Ca. Investigation of the properties of L_C' for $V_x=0$ to $V_x >> V_i$ shows it to be conservative [$\alpha' \leq 0.05$] with a limiting value when $V_x=0$ of approximately 0.03. Also in this limiting case, L_C' on the average is only slightly greater (< 10%) than it would be if one assumed V_x was identically zero, whereas for the conventional approach [$L_C = 2.92(s/\sqrt{n})$] it would be 78% larger. When the stakes are higher -- e.g., CI's or L_C's for $\alpha=0.01$ -- the contrast becomes even greater. In effect, we have used our knowledge of V_i to exclude very small values for estimated total σ, and gained smaller CI's, L_C's, and detection limits in return.

3.1.10 effects of rounding and truncation. Premature rounding of experimental data distorts its error distribution, resulting in erroneous conclusions regarding the shape of the distribution, its parameters [mean, variance], and results of statistical tests (e.g., detection decisions, quality of fit) and confidence intervals. The most obvious distortion is that an inherently continuous distribution is made discrete; the effect is analogous to "discretization noise" which is often found with multichannel and multidetector array techniques involving windows in time, space, energy, wavelength, etc. (56). The tolerable degree of rounding depends on the distribution. For normally distributed data, there is about a 10% chance of finding results within $\sigma/8$ of the mean. Scale divisions much smaller than $\sigma/4$ are therefore required if one is to avoid false coincidences, and fits that are "too good", etc. In fact, clues to excessive rounding or truncation may be found in χ^2 or F statistics which are unusually small, or in pdf's exhibiting unexpected deviations from normality (57). Abnormality is noted also by Cheeseman and Wilson for constrained balance-point measurements, such as the galvanometer needle which is physically confined to non-negative scale readings (36). The importance of these considerations for databases incorporating low-level results is discussed in (34).

3.1.11 the evaluation process [data reduction: fitting]. The data evaluation process [EP] is an integral part of the CMP, and as such it helps define σ_o and the critical level. It is perhaps obvious then that L_C, CI's, and the detection limit will differ for the very same experimental data, depending on the EP applied. A simple illustration is found in the fitting of spectral or chromatographic peaks. One may use the peak height as the quantitative signal measure, or a model-independent peak area may be used, or a more sophisticated technique such as linear or non-linear least squares may be employed to estimate the peak size according to a selected functional model such as a Gaussian or skewed Gaussian (58). The point is that without explicit specification of the entire CMP, including the EP employed, the detection characteristics of the measurement process are undefined. Because of this, a slight problem occurs when the EP is given as a "black box", or algorithm whose characteristics are unclear. (This issue, including the common availability of executable software

without source code, will be treated further in the discussion of detection limits in the next section.)

When the EP comprises linear computations (linear in the observations) such as simple differences, y - B, or linear least squares or linear multivariate computations, initial normality (of the observations y) is preserved for the estimated quantities. Non-linear computations, such as arise commonly in iterative model selection and peak search routines, produce estimated parameters having non-normal distributions (59). Caution is in order, in those cases, in applying "normal" values of test statistics to calculate L_c and CI's. (Other factors to consider are the extent of non-linearity, the level of confidence or significance [1-α], and the robustness of the statistic in question.)

Finally, it should be noted that an erroneous model will give erroneous results. This seeming truism is important because models which pass statistical tests [e.g., χ^2 test of fit] are consistent with the data but not necessarily correct. Because of multicollinearity, model error may go undetected, while producing significant bias in the results (48).

3.1.12 calibration error. A number of different approaches may be taken to incorporate the uncertainty in the calibration factor A into the critical level. To illustrate, let us consider the simplest functional relation for the Evaluation Process:

$$\hat{x} = (y - \hat{B})/\hat{A} = \hat{S}/\hat{A} \qquad (3)$$

Unfortunately, this is already a non-linear relation, so we cannot expect $\hat{x}$ to be normally distributed. If the relative error in A is small (e.g., < 10%) its influence on L_c is likewise small, and deviations from normality are minimal. If the relative uncertainty in A is not necessarily small, or if it includes possible systematic error, a straightforward approach is to use the lower bound for A to calculate an upper bound for L_c (here x_c) which can be used to make conservative detection decisions [$\alpha \leq 0.05$]. (Incorporation of bounds for systematic error is discussed more fully in the section on detection limits.)

Error propagation from the fitting of a calibration curve can be used to treat detection and interval estimation (almost) rigorously provided the model is correct -- a caution being that the intercept-B may not represent the blank-B (60,61). An interesting alternative is to estimate σ_o and the corresponding detection characteristics directly for $\hat{x}$ [i.e., in units of concentration] by full replication of the CMP at the levels of concern, observing y, B, and A for each replicate. (This is the "paired comparison" concept extended to calibration, where a blank and standard is run for every sample.) The statistical properties of the observed $\hat{x}$ distribution can then be used to directly calculate x_c [as ts_o, if A-variation is not too great] and estimate the detection limit. An added benefit of this scheme is that direct observation of the blank decouples it from the calibration curve fitting process, so that an assumed straight line model [constant sensitivity A] can be tested by fitting a line [S=Ax+e] through the origin (62). For valid conclusions, of course, due attention must be given to interference and matrix effects on both

parameters, B and A. An illustration of an observed $\hat{x}$ distribution for the null hypothesis [x = 0] is shown for ^{131}I in (44).

3.1.13 reporting of low-level data. Problems associated with data rounding and truncation extend to the reporting of final results. Also, just as in the case of the data evaluation step of the CMP, reporting must be treated as an integral part of the overall CMP. Bias and information-loss are the prime considerations. At the lower extreme, where x = 0 [null hypothesis] suppression of negative estimates forces a positive bias, on the average. Other biases arise when all non-detected results are reported as zero or as equal to (or less than) the detection limit. The difficulties are evident as soon as one attempts to: develop a database comprising large amounts of low-level data (34); to compute temporal or spatial averages for higher order detection decisions (28); or to compute average concentrations across different materials as in the USDIET-1 exercise (63). This last example illustrates the point: composite samples of the U.S diet were prepared for measurement of a broad range of essential and toxic chemical constituents, including the trace element Se. Comparison of the result for Se in the composite sample [128 μg/g] with the weighted average from the large number of individual contributing foods [100 μg/g], showed a significant negative bias for the latter. This was a result of setting all "trace" observations (defined as those below a quantification limit, L_Q) to zero. Adjusting these upward to $L_Q/2$ led to an improvement [110 μg/g], but negative bias was still apparent. These kinds of problems can be completely circumvented if concentration estimates, even if negative, are always reported together with their uncertainties (64, 65). Detection decisions can be made by comparison with L_C, and upper limits may be given as $\hat{x} + ts/\sqrt{n}$.

3.1.14 thresholds. The threshold for discriminating "real signals" from blanks may be set in various ways. The only way that is consistent with the relation between confidence intervals and significance tests is the one described, $L_C = ts$. Other techniques include the use of a constant multiplier ks or $k\sigma$, with k = 3 a popular choice; and use of a fixed threshold signal or concentration, such as 1 mV or 2 ng. A drawback of these alternative techniques is that they seldom recognize the existence or magnitude of the α-error, which, however, does exist, and which will take on varying values depending on the number of degrees of freedom or the magnitude of the fixed threshold in comparison to σ_o. For confidence intervals α is conventionally taken as 0.05, so there seems little justification for depressing it by a factor of forty to 0.0013 (corresponding to 3σ) for detection decisions [σ-known], or by more than a factor of ten (corresponding to 3s) for 20 degrees of freedom. Instruments having hardware or software discriminators may have resulting dead zones which correspond to vanishingly small α's. In one case recently, the threshold was set so high that α was beyond the range of any of the statistical tables, $L_C/\sigma_o \approx 34$; in fact the threshold was so high that the critical level exceeded even the conventional limit of quantification -- i.e., the RSD at L_C was but 5% (61, p. 76, case-e)! Such high and varying thresholds for detection decisions lead both to needless confusion and to measurement processes operating far short of their inherent capabilities.

3.2 Matters Concerning the Error of the Second Kind, and the Analyte Detection Limit. The concern in the preceding section was the validity of detection decisions, based on comparisons of experimental outcomes [estimated signals or concentrations] with appropriate critical levels or decision thresholds. Here, we turn to issues concerning the inherent detection capability of the CMP in question -- that is, the true signals or concentrations which can be detected with, for example, a 95% probability [β=0.05], given the critical level (or equivalently α) to be used for testing observed results. L_D is thus tied intimately to β, and to α or L_C. Although a significance test may be performed with no consideration of H_A or the detection limit, L_D is ambiguous without the specification of β and α (or L_C). That is, for a given L_D there is an infinite set of possible α, β pairs. Passing a significance test -- e.g., $\hat{x} \leq x_C$ -- is commonly said to mean "acceptance" of the null hypothesis -- i.e., $x = 0$. This is unfortunate terminology, for only consistency with the null hypothesis has been demonstrated. "Proof" of the null hypothesis (within certain fuzzy bounds) demands attention to all possible alternative hypotheses H_A; that is the test in use must be sufficiently powerful to "detect" [$\beta \leq 0.05$, given $\alpha = 0.05$] H_A. A major reason for interest in detection limits is thus to allow us to select or design a measurement process having the capacity to detect signals or analytes at prescribed levels of importance. An overview of selected technical issues follows.

3.2.1 ignorance of the error of the second kind (β). False negatives occur whether their existence is recognized or not. The common practice of making detection decisions at the so-called detection limit, or LOD, etc., has the effect of setting $L_D = L_C$, with the result that β = 50% -- equivalent to the proverbial flip of the coin. With a σ-coefficient of 3, α may be as small as 0.0013, resulting in an imbalance [β/α] of a factor of nearly 400! Ignorance of this matter makes possible inadvertent or even intentional misrepresentation of detection capability. For example, the subtle trade-off between α and β could be employed to avoid penalties for false positives associated with an inadequately controlled blank.

3.2.2 lower and upper detection limits. For certain types of chemical measurements there are dual null hypotheses and consequently dual L_C's and L_D's for concentrations differing from these null levels. Examples are found where a lower limit is set by background noise, and an upper limit, by some type of maximum signal limitation such as instrumental detector saturation. A dual illustration is shown in Fig. 7 for two exponential phenomena, radioactive decay and radiation absorption. In each case the lower L_D is given by the smallest detectable difference from a comparator (zero age standard or blank solution), and the upper L_D is given by the smallest detectable difference from an infinitely old sample (no net emitted radioactivity) or an infinitely absorbing sample (no net transmitted radiation.

3.2.3 the α - β connection: OC and ROC curves, and detection power. A convenient way to visualize the relationship between false positive [α] and false negative [β] errors and the normalized difference [d] between the means of two populations for a given statistical test has come to us from signal detection theory [9].

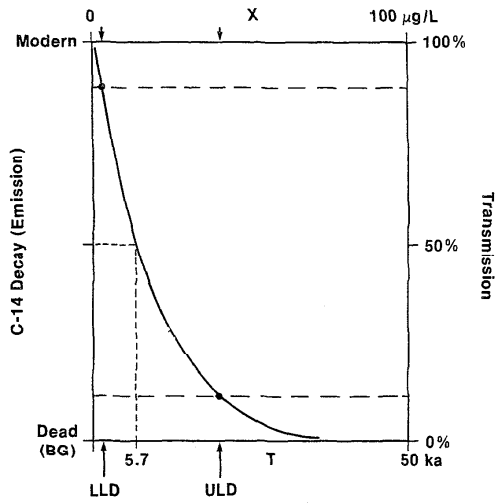

Fig. 7. Lower and Upper Detection Limits. When a measurement process has both minimum and maximum signal bounds, as in radioactive decay and optical absorption spectrometry, LLD and ULD must both be considered. Dashed line signal lower and upper detection limits map onto the age and concentration lower and upper limits (arrows) via the exponential function.

In this theory the Receiver Operating Characteristic [ROC] curve for a given test traces the relationship between the true positive probability [1-β] and the false positive probability [α] for a given mean normalized signal difference d. Another representation of the relationship, denoted the Power Curve in statistics, is the curve which traces the relation between Detection Power -- which is synonymous which the true positive probability [1-β] -- and the difference d, for a given value for α. (The complementary relation: β vs d, given α, is described in statistics as the Operating Characteristic [OC] curve.) Fig. 8 shows the normal ROC curve for d = 3.29 in units of σ_o [i.e., the detection limit], and the power curve for α = 0.05. The former [ROC] representation is the more convenient for the comparison of tests and the selection of alternative α, β pairs, for a given difference in population means. For this reason, it is used to compare the diagnostic power of alternative tests in clinical chemistry, where there are two discrete populations [d - fixed] (66). It may be useful also for examining the value of a test or the selection of an "optimal" α, β pair in a regulatory setting where, for example, a specified difference [d_R] is of concern. Also, if the sources or identities of chemical species are characterized by unique element or isotope ratios, an ROC curve could be used to represent the discriminating power of selected measurement techniques. The second [Power curve] representation is more appropriate when one is interested in the detection power as a function of (net) signal level or concentration. Thus, it is clear from the curve that the power is but 50% when d = 1.645. A second scale on the abscissa makes it convenient in this representation to see the relation in units of concentration.

OC and power curves are regularly used in the evaluation of statistical tests (67,68). Similarly, one finds ROC curves employed in medicine and psychology [12, discussion & references in Chapt. 25]. They appear to be little used in Analytical Chemistry, though Liteanu and Rica have proposed the use of different two dimensional projections of the three dimensional relationship [α, β, d] as representations of the "detection characteristic" (8).

3.2.4 power of the t-test. The three dimensional relationship described above is expanded to four for Student's t, with the addition of the number of degrees of freedom. If we restrict our attention to the detection limit, by fixing α and β both to 0.05, the remaining two dimensions can be viewed as a curve, d vs df -- i.e., the detection limit (in units of σ_o) as a function of the number of degrees of freedom, where $t_{1-\alpha}$ is used for making detection decisions. In this case, the value of d is determined by requiring a 95% probability (1-β) that the estimated net signal divided by its estimated standard deviation [$(y-B)/s_o$] will exceed the critical level for Student's t. This ratio is called the non-central t, with non-centrality parameter d, because it is displaced from zero by this amount. The net signal detection limit is given by $d\sigma_o$. Another important application of the non-central t distribution has been to test the validity of presumed detection limits, for example, in connection with medical diagnostic devices (69).

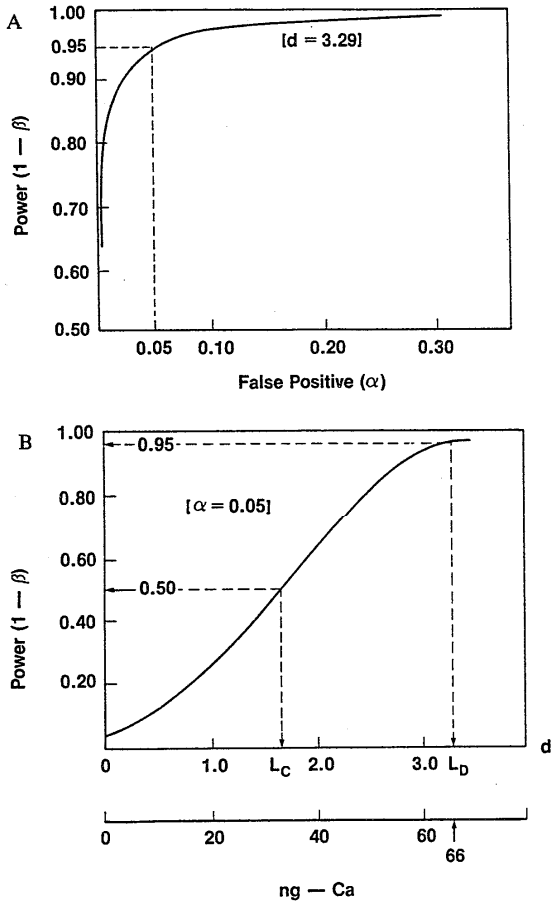

Fig. 8. Detection Power. ROC and Power (or OC) curves yield a graphical display of the relations among detection limits (detectable differences, d), and errors of the first (α) and second (β) kinds. Fig. 8A is the ROC curve corresponding to two normal populations differing by L_D -- i.e., the separation equals 3.29 σ_o, and the curve passes through the point $\alpha=\beta=0.05$. Fig. 8B is the corresponding power curve, where now α is fixed, and the power of the test is given as a function of the normalized distance d. The lower abscissa shows the equivalent concentration scale for a hypothetical measurement process for Ca, where σ_o equals 20 ng, and the detection limit is 66 ng.

When df is very large, d is simply 2z or 3.29. For fewer degrees of freedom, 2t yields a conservative estimate, but a still better estimate derives from the following expression (68, p. 252):

$$d \approx 3.29\ (1 + 0.71/df) \qquad (4)$$

This formula is accurate to about 1% or better for df ≥ 8. For 4 - 7 degrees of freedom the correct values are 4.07, 3.87, 3.75, and 3.68. To illustrate, let us suppose that 5 paired y, B observations were made and the mean difference and estimated standard error were 1.8 ± 1.2 mV. The critical level for 4 degrees of freedom would be ts = (2.13)(1.2) = 2.6 mV, so the conclusion would be "not detected." The detection limit would be $d\sigma_o$ = $4.07\sigma_o$. Using s as an estimate for σ, we would estimate L_D as (4.07)(1.2)=4.9 mV.

3.2.5 uncertainties in detection limits. The previous example raises an extremely important point, namely, that unless σ_o is known without error, the detection limit cannot be exactly known. This is in contrast with the critical level, which can always be explicitly calculated from Student's t and the estimated standard error. We can, however, derive a confidence interval for L_D from the bounds for σ, given s and df. For normally distributed errors these bounds can be derived from the χ^2 distribution. (s^2/σ^2 is distributed as χ^2/df.) One finds, for example, that at least 13 replicates are necessary to obtain s within 50% of the true σ (90% confidence level).

For practical application of detection limits -- e.g., in meeting a research or regulatory requirement -- a "safer" procedure is to quote the upper limit for L_D. This in effect casts the uncertainty onto β, in that a specific value (rather than a range) can be given for the detection limit, but with the proviso that $\beta \leq 0.05$ (with 95% confidence). A straightforward, conservative treatment for detection decisions and detection limits when s is estimated from replication is thus: to use L_C = ts (α=0.05) for detection decisions; and to use $L_D = 2ts(\sigma/s)_M$ (α=0.05, $\beta \leq 0.05$) for detection limits. From the brief table of the relevant quantities which follows, we see for example with n = 10, L_C = 1.83 s, and L_D=2(1.83)(1.65)s = 6.04 s (to be compared with 3.29 σ for df = ∞.) [Table II; from Ref. 28, p. 80].

Table II. L_D Estimation by Replication: Student's-t and (σ/s) - Bounds vs Number of Observations

No. of replicates:	5	10	13	20	120	∞
Student's-t:	2.13	1.83	1.78	1.73	1.66	1.645
σ_{UL}/s:	2.37	1.65	1.51	1.37	1.12	1.000

A second source of uncertainty is associated with the quantities comprising the overall calibration factor A, such as recovery, instrumental detection efficiency, matrix absorption or scattering, etc. If A is determined as a random variable each time x (concentration) is estimated, then there is no problem; its random error is automatically taken into account through error

propagation or replication when σ_x is estimated. If the same estimate for the calibration factor is repeatedly used, its random error has become a bias, and the bounds (confidence interval) for this bias combined with other possible sources of A-bias produce an uncertainty interval in the concentration detection limit. The recommended approach, again taking into account the practical applications for detection limits, is to transfer the uncertainty to β by taking the upper limit, S_D/A_m, as the concentration detection limit with $\beta \leq 0.05$.

3.2.6 uncertainty bounds (systematic error) for the blank. If the possibility of significant bias in the estimated value for the blank is not taken into account, the resultant detection decisions and limits may be much too optimistic. An upper limit for this bias component can be incorporated into S_C and S_D estimation just as it is in total uncertainty interval estimation, by extending the random uncertainty (confidence) limit by the upper bound for bias Δ_M. Thus, S_C becomes $S_C' = 1.645\ \sigma_o + \Delta_M$, and $S_D' = 2\ S_C'$. The detection limit increases therefore by $2\ \Delta_M$. The rationale for this procedure is indicated in Fig. 9. For a number of measurement disciplines, experience dictates reasonable values for **relative** limits for blank and calibration factor bias [$\hat{\Delta}_B, \hat{\Delta}_A$]. Default values of 5% and 10%, respectively, have been suggested (28) and tentatively confirmed (57) for radioactivity monitoring, for example. In this case, $S_C' = 1.645\ \sigma_o + 0.05\ B = S_C + 0.05\ B$, and $x_D' = 1.1\ (2\ S_C')/A = 1.1\ x_D + 0.11$ BEA, where $x_D = 3.29\ \sigma_o/A$ and BEA is the background equivalent activity. Thus for paired measurements with B = 500 counts ($\sigma_o = \sqrt{2B} = 31.6$ counts), and A = 5.0 count/pCi, $S_C' = 52 + 25 = 77$ counts, and $x_D' = 22.9 + 11.0 = 33.9$ pCi. Clearly, detection in this case is neither fully statistical nor fully non-statistical. Balancing the limits imposed by the statistics of signal detection with those derived from our knowledge (or ignorance) of the measurement process is essential for meaningful decision making. Historical use of a multiple of the blank is perhaps more readily understood also, through the formal incorporation of the term 0.11 BEA.

3.2.7 optimization and iteration: figure of merit. Optimal detection limits are sometimes treated through the maximization of some sort of figure of merit (FOM) such as $S/\sqrt{B}$, etc. Simplistic FOM's tend to ignore complex dependence of L_D's on measurement conditions, systematic error components, and the explicit nature of the sample. As shown in (35) for example the variation of radioactivity detection limits with counting time may range from t^{-1} to t^{+1}. Since detection limits may be sample-dependent, because of interference and matrix effects, iterative estimation of the detection limit is sometimes required. Changes in the measurement process may also be necessary if such sample dependence forces the actual detection limit above the corresponding regulatory limit.

3.2.8 multicomponent detection limits. When one leaves the realm of "simple" y - B net signal estimation, modeling and linear or non-linear least squares computations are generally required for component estimation. For the linear multicomponent model it is possible to estimate the detection limit as a closed expression, provided that all interfering analytes are included and the errors (variance), constant. Weighted least squares calculations involving Poisson or other concentration-dependent statistical

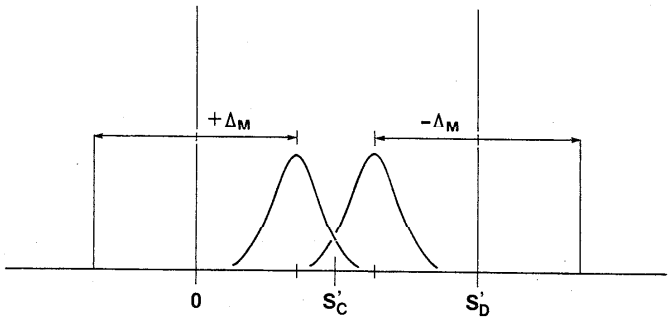

Fig. 9. Effect of Bias on Detection Limits. Allowance for bounds for bias [Δ_M] increases the critical level by Δ_M, and the detection limit by twice that amount (simple detection), taking the sign for the uncompensated bias as unknown. α and β are now inequalities -- ie, α, $\beta \leq 0.05$.

weights require an iterative L_D estimate (13,61). A relatively simple inequality expression for L_D can be given, however, by using the results of a linear model fit, where the component of interest is present at or above its detection limit (28). That is, $x_D \leq 3.29\ \sigma_x$ where σ_x is the standard error for the least squares estimate of the component in question. The basis for this inequality is that the standard error for any specific component will be approximately constant or increase with its increasing concentration in a mixture, all else remaining equal. Special multicomponent problems associated with selected types of pattern matching and multiple independent decisions, as in the detection of several isolated spectral or chromatographic peaks, will be treated in the next section.

3.2.9 effects of gradually changing distribution functions and/or σ's. For $\alpha = \beta$, normally distributed errors with constant, known variance, and simple paired estimates (y - B), $L_D/L_C=2$. This simple ratio does not obtain, however, whenever any of these conditions are not fulfilled. The matter of high thresholds ($\alpha << 1$) has already been noted. In the example discussed in the last paragraph of the preceding section, α was vanishingly small when L_C was set to $3\frac{1}{4}\sigma_o$. The ratio L_D/L_C in this case was 1.09 (61). Similarly, the ratio is unity when the conceptual difference between L_C and L_D is overlooked, such that $\beta = 0.50$.

Changing cdf's is another matter. Because the overall detection process in effect relates to the discrimination of net signals at the detection limit from null signals, one is faced with the possibility of two different distributions at the two levels, S=0 and $S=S_D$. This problem does not arise in making detection decisions, however, for S_C depends only on σ_o, the standard deviation of S when its true (mean) value is zero. Two cases will be considered, a) the Poisson counting distribution, which changes shape and (relative) discreteness with increasing signal level; and b) the normal distribution, where σ increases with concentration, a common occurrence in analytical chemistry. A third case of some importance for environmental measurements is the distributional perturbation which occurs as one adds normal measurement errors to log-normal blank variations.

The Poisson distribution is decidely asymmetric and discrete (in a relative sense) at the lowest levels. In fact, when the expected (mean) value of the Poisson parameter -- in this case, the blank -- is smaller than 0.05 counts, the critical level (y_C, gross counts) equals zero. (This quantity is necessarily an integer for the Poisson distribution, and α must be treated as an inequality [$\alpha \leq 0.05$].) The detection limit (y_D, gross counts) then equals 3.00 (not necessarily an integer), so y_D/y_C in this case is infinite. With increasing signal level (counts) the Poisson distribution approaches normality, so "the usual equations apply," and L_D/L_C approaches 2.0. For B = 1.0, $L_D/L_C \approx 3.4$; for $B \geq 5$, $L_D \approx 2.7+2 \cdot L_C$, with $L_C = 1.645\ \sqrt{B}$ is a good approximation. See (28) for a more extensive treatment of extreme, low-level counting statistics.

Poisson σ's increase with $\sqrt{(S+B)}$. A linear increase of σ with concentration [$\sigma(y) = \sigma_B + mS$], however, is common for many analytical methods. Since $S_C = z\sigma_o$ and $S_D = S_C + z\sigma_D$, the increase in σ in passing from S=0 to $S=S_D$ means that $S_D > 2\ S_C$. A closed expression, however, may be derived using the above linear model. (To

simplify the notation in the remainder of this paragraph, S_C and S_D will be represented as C and D respectively.) Let us consider first the case where a precise estimate $\tilde{B}$ is available for B.

Standard deviations are given by:

σ, assumed variation: $\sigma_y = \sigma(S+B) = \sigma_B + mS$

σ, null signal [S = 0]: $\sigma_o = \sigma(B-\tilde{B}) \approx \sigma_B$

σ, detection limit [S = D]: $\sigma_D = \sigma(D+B-\tilde{B}) \approx \sigma(D+B) = \sigma_B + mD$

For normal random errors and $\alpha = \beta = 0.05$, the critical level and detection limit are defined as:

$$C = z\sigma_o \approx z\sigma_B = 1.645\ \sigma_B$$
$$D = C + z\sigma_D \approx C + z(\sigma_B + mD)$$

The last equation may be solved for D:

$$D = 2C/(1-zm)$$

For $\alpha=\beta=0.05$, i.e., z=1.645, an important conclusion emerges: the **detection limit does not exist for** $m > 1/z = 0.61$. This may be academic, however, since so large a slope is unlikely for any reasonable analytical method. A slope of 10%, however, would result in D/C = 2.39. To illustrate, let us take the blank standard deviation for the measurement of toluene in air, by a fully specified method of sampling and gas chromatographic analysis, to be 0.21 μg/L. The critical level for detection decisions, assuming normality, would then be 1.645 (0.21) = 0.34 μg/L. The corresponding detection limit would be $2.39(0.34_5) = 0.83$ μg/L.

For the general case, where B is estimated from n replicates, the algebra is only slightly more complicated. The variance of the estimated net signal is now given by:

$$V_S = (\sigma_B + mS)^2 + \sigma_B^2/n = V_o + mS\ (2\sigma_B + mS)$$

thus, defining $\eta=(n+1)/n$:

$$V_o = \sigma_o^2 = \sigma_B^2\ \eta \quad \text{and} \quad V_D = \sigma_D^2 = V_o + mD\ (2\sigma_B + mD)$$

From these relations and the definitions for C [$C = z\sigma_o$] and D[$D=C+z\sigma_D$], it is relatively straightforward to show that:

$$D = 2C\ [1 + zm/\sqrt{\eta}]/[1 - (zm)^2] \text{ and } \sigma_D/\sigma_o = D/C - 1$$

Thus for $\eta=2$ (paired comparison), and m and z as before (0.1, 1.645), D/C = 2.29 and $\sigma_D/\sigma_o = 1.29$. The asymptotic result (D=2C) follows of course when the slope m is negligible.

3.2.10 black boxes and hidden algorithms. With the advent of "user friendly" (and proprietary) software and automated data reduction and even automated instrument systems which yield final results only, a cautionary note must be sounded. That is, when the computational scheme is not fully and explicitly described, and when the software is not exhaustively studied and tested, erroneous results may emerge. Worse still, there may be no way of recognizing such results as erroneous, particularly if the

instrumental system is designed in such a manner that the raw experimental data cannot be retrieved for alternative methods of computation. It is inappropriate in this chapter to document the problems arising, but it may be helpful to glimpse at their nature. Inept or unlucky programming and inaccurate stored parameters will always cause difficulties, but this of course is not restricted to the domain of low-level measurement. Problems of special concern for reliable and efficient detection which have come to the attention of this author include: a) Thresholds which are automatically set so high that detection power is seriously eroded; b) Algorithms (and component models) which are data dependent. This is especially a problem when peaks are marginally discernable, with peak estimation algorithms switching rules depending on the magnitude or apparent presence of a peak; c) estimation and search routines based on inadequate models or inadequately accounting for the effects of non-linear estimation; d) decision or detection algorithms for which assumptions or parameters used are unclear (and possibly incorrect); e) inaccessibility of raw data, especially when peaks are not found, and the consequent inability to investigate extra sources of variability or errors in assumptions, models or data.

Peak search or "model search" and more generally optimization routines that are sometimes heuristic and operate strictly in an empirical fashion on the data at hand, deserve another comment. That is, at the lowest levels and especially at the $S = 0$ extreme [null hypothesis] such generally non-linear routines may provide no S estimates, especially when negative, thus producing biased and skewed S distributions. Once a peak or model is automatically chosen from the noisy data, the algorithm switches, frequently to a linear estimation algorithm. The problem is that the switch point varies, being noise controlled; also the estimation algorithm seldom gets to operate when the null hypothesis [S=0] is true. σ_o is not obtained, and the normal distribution hypothesis testing apparatus cannot be applied at the lowest signal levels. Perhaps this is why the international gamma ray peak detection exercise organized by the International Atomic Energy Agency found the "visual" method of peak detection more successful than all others, including the most sophisticated computer based schemes (58).

The "model search" issue is more profound. In multicomponent chemical analysis, optimal models for estimation (number and nature of components) are often chosen automatically and empirically, for example by applying iterative, non-linear optimization routines, and quite frequently non-negativity constraints. Such automatic chemical model building, accomplished by suppressing (often legitimate) negative estimates, deserves careful scrutiny. It may be even more misleading than zero suppression with simple measurements, especially when noise and multicollinearity are large.

Illustrations of some of these limitations, which are unique for low-level data and therefore meaningful detection limits, may be found in references 35, 57, and 70. Fig. 2 in the first reference illustrates an extreme, yet not uncommon problem; quite visible spectral peaks have failed to be detected by the software.

3.2.11 quality. One solution for inadequate or incorrect approaches to detection -- including control of both false positives and false negatives -- is the incorporation of known and

blind standard reference samples and reference simulated data. Such means for control are well established for trace analysis, but they have rarely been brought to bear on the detection problem. Interlaboratory low-level test data, though quite rare, have proven most informative (58,70). Direct validation and control of α and β errors should be made routinely with blind interlaboratory samples and/or data representative of blanks and samples at or near detection (or regulatory) limits, respectively. Evaluation of sets of results via ROC curves could, in turn, be quite fruitful, for the quality of the low-level measurements would be reflected in the loci of the ROC curves, *independent of the particular decision rules employed.*

3.3 Discrimination Limits, Multiple Detection Decisions, and Patterns. When the null hypothesis is defined as zero analyte concentration beyond the blank, or zero signal above the baseline or background, it is appropriate to refer to an analyte (or signal) detection process. In many practical cases, however, it is interesting to consider the ability to discriminate concentrations from a fixed non-zero reference level, or discriminate patterns or "chemical fingerprints" from a reference pattern. Multiple hypothesis testing decisions form a natural link between these two types of discrimination, and it becomes clear that both fixed level "recognition" and chemical pattern recognition fall under the same statistical frame-work as zero level analyte detection. Both aspects of Qualitative Analysis (detection, identification) share the same probabilistic foundation, including hypothesis specification [H_o, H_A], decision criteria, and type I [α] and type II [β] errors [Ref. 8; pp. 233, 239]. In all cases, it is extremely important to recognize that the respective discrimination or detection limits characterize the measurement process, not a particular result. (As always, results are tested by comparison to the corresponding critical level.) Our objective is to evaluate the intrinsic capabilities of CMP's, often shaping these capabilities to meet specific practical or research needs.

3.3.1 lower and upper regulatory limits: balancing risks and costs. We have noted that detection limits dictated by regulatory concerns have been surrounded by considerable confusion, discrepant statistical and ad hoc formulations, ignorance, and even mild deception. The apparent deception is related to the lack of general understanding or agreement concerning the appropriate nature and magnitude of the error of the second kind (β, false negative). By ignoring its presence, whether intentional or not, those who must meet regulatory demands generate a β/α imbalance where, at 50%, false negatives may exceed false positives by nearly a factor of 400. One justification is that identically zero concentrations cannot exist anyway, and very small concentrations cannot be effectively distinguished from the blank. A related, very important observation is that small non-zero concentrations will be "detected" on occasion, necessarily more frequently than the false positive constraint [α] placed on the blank. To reduce the penalty which might be associated with the occasional detection of such small concentrations, it is of course helpful to reduce α for the blank still further -- but this should be done openly, not by subterfuge.

To meet such legitimate concerns, while at the same time keeping an open, realistic, and balanced view of false positives and negatives, we recommend the substitution of lower and upper regulatory limits -- whose difference is the discrimination limit (Δ_D) -- in place of the null limit [zero] and the single, analyte detection limit. To illustrate the suggested approach, Fig. 2 has been modified in Fig. 10 to indicate a non-zero lower limit (L_Q), and an upper limit (L_D). As before, the upper limit in the societal or regulatory setting would be established at such a level that the concentration or event of concern would be reliably detected [β=0.05] when its net "cost" to society crossed the limit of acceptability. The lower limit from which the upper limit must be reliably discriminated, is new: its level is established such that the penalty for "detecting" a very small concentration is likewise acceptable. Such penalties can be quite real, especially in terms of intangibles, such as public alarm (71), or indirect long-term negative perceptions affecting the business of a regulatee. In Fig. 10 this concept is presented, again in the context of earthquake detection, with the aid of hypothetical positive and negative cost differentials which would define the "trigger points" for L_Q and L_D.

It is important that the (regulatory) level-setting process for these limits be decoupled from their estimation from the characteristics of the measurement process. The former is a sociopolitical matter involving complex risk assessment issues (4), whilst the latter lies in the domain of the scientist. The scientific responsibility is met once the discriminable limits lie within those desired by society. Note that the discrimination limit Δ_D is here defined as the difference L_D-L_Q such that α, β each equal 0.05. It is interesting next to consider precision requirements, e.g., at the upper limit, as compared to those for the conventional detection limit. Taking L_Q to be 50% of L_D, the relative standard deviation at this L_D would be about 15%, in contrast with 30% for the conventional detection limit. (The change is entirely due to the introduction of the non-zero L_Q; the magnitude of L_D is unchanged.) The precision (RSD at L_D) would be "quantitative" (10%), once L_Q equals 2/3 of L_D. **Quite possibly the (subconscious) need for such discrimination capability is the underlying motivation of those who call for abandoning detection limits and hypothesis testing in favor of "quantitative" measurements.**

The discrimination limit as depicted in Fig. 10 has two other important applications, one in business and one in science. In business matters involving trade or regulation, one may face the task of "proving" the product or waste stream level exceeds or does not exceed some prescribed value, such as the (upper) regulatory limit. Because of measurement error, the ability to accomplish this is limited, and in fact it is set by the size of the discrimination limit. Balancing of costs will again generally fix the magnitude of Δ_D. Penalties will likely increase with greater apparent departures from specifications; and the ability to defend departures as small, or attack departures as large depends upon the producer's or consumer's discrimination limit. The discrimination limit, hence precision of analysis, can only be improved with increased analytical costs. In the socioeconomic arena decision

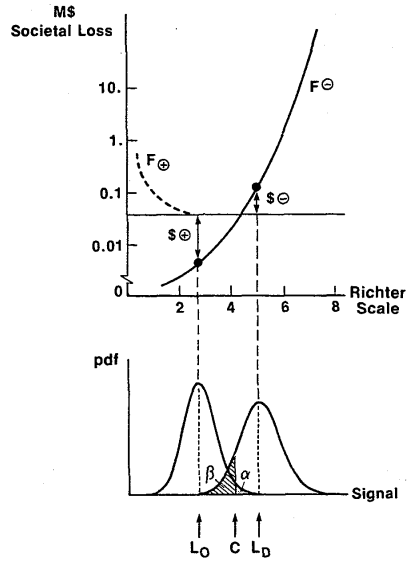

Fig. 10. Discrimination Limits. Curve F^- represents the loss to society as a function of earthquake magnitude; F^+ represents the cost of avoidance (evacuation, etc.), the dashed portion simulating *indirect* costs associated with false alarms -- eg, mental anguish, damaged credibility, lawsuits, etc. Points of imbalance between F^+ and F^- which exceed what is acceptable to society are taken as lower and upper regulatory limits, which must be matched by corresponding lower (L_o) and upper (L_D) measurement limits whose difference is the Discrimination Limit (Δ_D). A non-zero lower limit forces an improved precision requirement in comparison to the "simple" L_D of Fig. 2.

theory may be helpful for deriving the appropriate balance between penalties and analytical costs (therefore the requisite Δ_D), particularly taking into consideration which party has the burden of proof. (Note that the "cost" *differential* associated with the *burden of proof* is equivalent to the size of the [measurement] *dead zone* around the regulatory or specification limit -- i.e., the sum of the "producer's" and "consumer's" discrimination limits.)

The scientific application involves "**identification**" in its simplest sense. That is, if L_0 and L_D are treated as unique or identifying concentrations or isotope ratios, or characteristic energies or wavelengths, etc., then the measurement process must be designed so that Δ_D is sufficiently small to distinguish between these two classes. In analogy with the detection power $(1-\beta)$ characterizing the detection limit (given α), one finds the power associated with Δ_D described as "discriminatory power" (12, p. 517) or "resolving power." This univariate, statistical approach to identification shares much in common with detection. For example, OC and ROC curves are just as appropriate for balancing false positives and negatives, and for comparing capabilities of alternative measurement (and computational) techniques. In addition, the difference between **design** of the measurement process to achieve a given detection or identification capability and **outcome** (specific result) is still manifest in an uncertain region -- i.e., results falling within the RUD[region of uncertain detection] or RUI [region of uncertain identification] **may** be detected or identified, respectively, by chance but this cannot be "assured" (α, $\beta = 0.05$) a priori. (See "multichannel identification, below, for further discussion.)

3.3.2 impurity detection. A special issue involving discrimination limits in analytical chemistry, having broad importance, is the detection of impurities or contamination. Conceptually, this can be treated as a direct outgrowth of the "identification" or discrimination of singular classes characterized by unique values of a continuous variable, as described in the preceding paragraph. In Fig. 11 class-0 and class-A are shown at separate unique (identifying) locations of a continuous measurement variable x_i. As depicted, the separation of these two classes far exceeds the discrimination limit Δ_D, so identification of a pure component (in this 2-component universe) will present no problem. If component-0 is contaminated by a small admixture of component-A, however, there exists a limit [Δ_D] below which a contaminated sample will be indistinguishable from the pure component-0. The **minimum detectable contamination** is numerically equal to Δ_D, when Δ_D is expressed relative to the class separation $(x_A - x_0)$ -- i.e., as a mole fraction or mixing ratio. (Note that "mixing" can occur as physical mixing of miscible chemical species, or it can arise from superposition of signals from different sources within the same detector.)

Two fundamental observations follow. First, class separability and impurity detection power degrade with increasing variance of the $\hat{x}_i$ distributions, which in turn, depends on the measurement precision and therefore the detection limits for the two components. This direct, and quantifiable coupling between pure component detection limits, component identification and resolution, and impurity detection is most important, though

scarcely surprising. Second, if one has sufficient knowledge of the "chemical universe" -- i.e., x_i locations for the entire population of H_A's -- then for any H_O of interest, one can deduce the maximum systematic error due to undetected contamination by estimating Δ_D for the "closest" impurity source. If this discrimination limit is unacceptable, redesign of the CMP is in order. Reliable estimation of systematic error bounds deriving from undetectable contamination, or undetectable model error is one of the important needs for accurate analytical results. Thoughtful consideration of the coupling between experimental design, and component detection and discrimination limits, supported by excellent scientific knowledge concerning the H_A universe offers one of the most reliable and objective solutions to this problem. An astute examination of these issues, emphasizing the universe of potential contaminants has been provided by Rogers (6). For simplicity, the discrimination problem was presented here in one dimension (one measurement variable). Multivariable detection and discrimination are obvious extensions, leading generally to increased detection and discrimination power, as one compares or "matches" unique multivariable patterns in place of characteristic values of a single variable. (See below.)

3.3.3 multiple detection decisions. If a number of detection or discrimination decisions are made in the course of a measurement, the overall probabilities of false positives and false negatives are accordingly altered. We consider two cases: first, where the individual tests are unrelated or "serial", and second, where "parallel" tests are made, as in pattern recognition. Independent, serial tests characterize the detection of isolated spectral peaks, as in multichannel gamma ray spectroscopy, as well as residuals following data analysis, and even replication experiments and control charts. In all of these cases, the overall probability of false positives and false negatives necessarily exceeds that for the individual peak detection (or outlier detection) test. For example, if a large gamma ray spectrum containing no actual radioactivity were scanned with the equivalent of, say, 50 detection decisions [$\alpha = 0.05$], there almost certainly [>92% chance] would be at least one false positive peak. Similar considerations apply to false negatives, so false alarms and missed radioactivity would be the consequence. (Ignoring this issue has led to some difficulties in the evaluation of low-level gamma ray spectra; see Ref. 28 for further discussion.) The solution is to follow the rules for combining probabilities; namely, adjusting the significance levels so that the overall probabilities of correct non-detection [$1-\alpha'$] and correct detection [$1-\beta'$] remain 95%. The probability that all decisions are correct is simply the continued product: $(1-\pi') = 0.95 = \Pi(1-\pi) = (1-\pi)^n$, where π represents α or β, and n, the equivalent number of tests per spectrum. Adjusted values for α and β are then given by (Eq. 5),

$$\alpha \text{ (or } \beta) = 1 - (0.95)^{1/n} \qquad (5)$$

If the total error level is to be held at 5% [α', β'] for a multitest experiment in which H_o is actually true 50 times and H_A, 3 times, then Eq. 5 gives adjusted values of $\alpha = 0.00103$, $\beta = 0.017$ with corresponding critical levels and detection limits of $3.1\ \sigma_o$

and 5.2 σ_o, respectively. Monitoring of mostly empty spectra thus provides justification for unequal α, β, and thus for $L_D/L_C < 2$.

3.3.4 multichannel identification. The linkage between detection and identification was brought up earlier, where "identification" was formulated in terms of the statistical estimation of the characteristic value (e.g., element concentration or ratio, gamma ray energy) for the identifying variable. Linear estimation was at least implied in that discussion, so that initially normal data would lead to normal (though possibly correlated) errors for the estimated results. For example, the frequency distribution of events (counts) along the energy axis [identifying variable] could be used to estimate the mean energy ("centroid") and its variance for a gamma ray peak, and the peak magnitude or "area" could be simultaneously estimated with a simple filter function to compensate for a linear baseline. The decision space is now two dimensional, so contours of the bivariate area (detection) - energy (identification) distribution would be used for significance testing. When multichannel data are intrinsically non-normal, or when they are subjected to non-linear operations as in certain peak search and peak fitting algorithms, normality is not preserved, so caution is in order in making detection decisions and in deriving confidence intervals.

"Non-statistical" identification is important in many facets of analytical science, where signal location or "identity bin" predetermines species identity. Detection and identification are then uncoupled, and any signal detected in the characteristic bin simultaneously conveys detection and identity. Classical analytical chemistry (e.g., gravimetry) relied heavily on this model, where unique chemical separations would guarantee identity. Modern instrumental or chromatographic methods similarly succeed when the resolving power (discrimination power) far exceeds the "density" of pure components along the informing variable.

3.3.5 pattern discrimination limits [multivariable identification]. We considered the discrimination of chemical components or classes earlier from a univariate perspective, including the paired comparison for a single alternative or contaminating component which would necessarily lie to one side of the null class (known component against which the sample is to be compared). Before considering discrimination with multiple chemical variables (compositional or spectral patterns), let us broaden the univariate problem to two-sided discrimination, since unlike analyte detection, characteristic or identifying variable values generally may be larger or smaller than that of H_o. The H_o discrimination limit test would then be 2-sided -- i.e., $z_C = 1.96$ instead of 1.645 for $\alpha = 0.05$. If a paired comparison of a test sample (unknown) with the control sample (known, H_o) falls within $\pm 1.96\ \sigma_o$, we then conclude that there is no statistically significant difference. This is not, however, proof that the patterns are the same; it is only a test of consistency. It is necessary, but not sufficient. To establish a real match, or "identification," we must demonstrate that the universe of alternative patterns will not match (statistically). Design of a measurement process for the successful identification of a particular chemical species or compositional state thus requires consideration of both α and β errors, as depicted in Fig. 12.

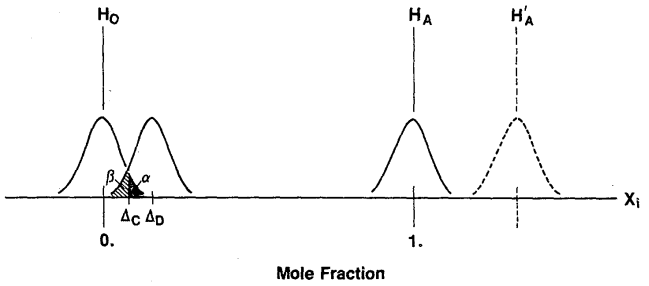

Fig. 11. Impurity Detection. Δ_D represents the minimum detectable concentration of substance-A [H_A] in the "null" substance [H_o]. The abscissa represents mole fraction or mixing ratio. Individual impurity detection limits would obtain for each impurity type, e.g., A' [H_A'].

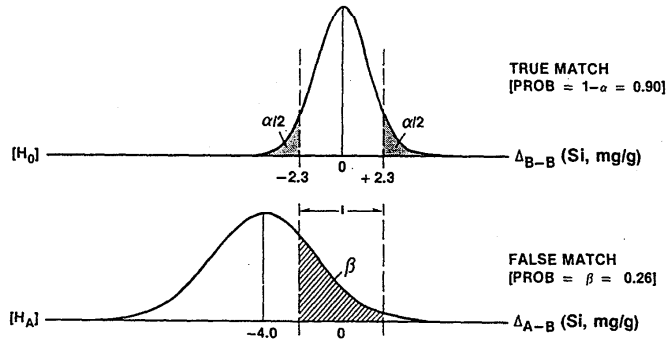

Fig. 12. Single Species Matching; Univariable Identification. For a given location on the abscissa [identifying variable: isotope ratio, X-ray energy,...], unique identification requires that none of the possible H_A's overlaps (probability β or less) the two-sided H_o window [I]. That is, all separations must exceed the corresponding discrimination limits. (From the design perspective, since identifying variable separations are generally fixed by Nature, we must design the CMP to achieve corresponding Δ_D's -- cf, Fig. 8A [ROC curve].) [Illustration constructed using $\alpha = 0.10$, $\Delta = -4.0$ mg/g, $\sigma_B = 1.0$ mg/g, and $\sigma_A = 2.6$ mg/g.]

It is a small step to take from univariable identification to multivariable or pattern matching. If we are concerned with just a single alternative pattern [A], but several (n) measured variables, then the consistency test requires that all n variables match statistically when the identity of test sample is the same as that of the control sample [B]. Combining probabilities as before, $(1-\alpha') = 0.95 = \Pi(1-\alpha_i) = (1-\alpha)^n$. Proof of identity, as before, includes consideration of sufficiency -- i.e., we require in addition that [A] not match (statistically) [B] simultaneously for all measured variables. Probabilities are combined a little differently in this case; the overall probability of an erroneous match is given by $\beta' = \Pi(\beta_i)$. The product is also taken over all n variables, whose individual β_i's will generally differ. Unless $\beta' \leq 0.05$, matching of patterns cannot establish identity. At the same time, it is this multiplicative feature, when individual β's are themselves small, that gives multivariable or pattern discrimination its enormous power.

To illustrate, let us consider matching of trace element patterns in two pure source materials, where the origin of one (control sample, B) is known, as is the composition of the possible alternative A. Given the characteristics of the measurement process and the compositions of the two known sources, we can tell a priori whether the sources are discriminable as indicated above. If not, the capability of an unknown test sample to match proves nothing. Absence of a match under these conditions, however, would deserve scrutiny; it could indicate either faulty measurements or faulty assumptions. Illustrative data are given in Table III.

Table III. Multivariable Identification

Input data for estimating the discriminability (identifiability) of particle emissions from steel plants A and B [a,b,c]

	H_o: B vs $\hat{B}$			H_A: A vs $\hat{B}$		
	Al	Si	Ca	Cr	Mn	Fe
Concentration (mg/g)						
steel-B	10	12	45	3.2	22	160
steel-A	13	8	70	3.3	16	120
σ	1.1	1.0	5.8	0.32	1.9	14
window [I ±]	4.00	3.63	21.1	1.16	6.91	50.9
distance [Δ]	3.0	-4.0	25.	0.10	-6.0	-40.
β	0.74	0.40	0.32	0.98	0.63	0.71
Δ/σ_o	1.93	-2.83	3.05	0.22	-2.23	-2.02

(a) Based on data from Ref. 72.
(b) Values of I and β are given for n=5.
(c) Fig. 13 depicts the windows [I] and variable separations [Δ].

Concentrations for six elements characterizing two steel aerosol samples (72) are given in the first two rows. Steel-B is taken as the control, and steel-A as the alternative source. H_0 is represented by the vector or pattern difference, $(\mathbf{x}_B - \hat{\mathbf{x}}_B)$; H_A, by $(\mathbf{x}_A - \hat{\mathbf{x}}_B)$. The last five rows of the table indicate, respectively: the standard deviations $[\sigma]$ for the elements in question, the matching intervals [I], the concentration differences $[\Delta]$ under H_A, the probability of false matches $[\beta]$, and the ratios of concentration differences $[\Delta]$ to the paired measurement standard deviations $[\sigma_o]$. $1-\beta$ and Δ/σ_o both serve as measures of individual element discriminating power. The quantity I is computed by requiring $1-\alpha'$ to be 0.95; for n=5, this means $\alpha = 0.0102$ or z_c (2-sided) = 2.57. (For 6-member patterns, z_c increases to 2.63.) Then $I = \pm z_c\sigma_o$, where $\sigma_o = \sigma\sqrt{2}$. Pattern differences $[\Delta]$, indicated by the open circles, are shown in comparison with matching intervals in Fig. 13.

For this example, pattern identifiability (H_o "provability") has been approached in two ways. First, β' has been calculated as the product of the individual β_i's, reflecting the series of individual element matching decisions. (For n = 5, omitting Ca, this product equals 0.13.) Second, the vector difference represented by H_A is examined through the use of the non-central χ^2 statistic, where $\Sigma(\Delta/\sigma_o)^2$ is the non-centrality parameter (73). In this second case the test of the vector match (i.e., H_o test) is carried out by comparing the sum of squares of the n observed normalized differences with the critical level for the central χ^2 for n - degrees of freedom. The rms value from the sum of squares -- $(\Delta/\sigma_o)_{rms}$ -- represents the multivariable generalization of the univariate normalized differences. It is a convenient single parameter measure (index) for the vector discrimination power $(1-\beta')$, as β' is uniquely determined by this quantity, given α' and the number of degrees of freedom.

Table IV gives results for the two types of test and several choices of element patterns. Important dual pattern identification conclusions follow: (a) Discrimination power (identifiability) differs according to the type of test, χ^2 being significantly better and becoming more so with increased dimensionality. (b) Optimal feature selection (e.g., for n=5) gives optimal discriminating power for the number of variables selected. (c) There exists an optimal number of dimensions (variables). The most powerful variable (here, Ca) is used for n=1; a second discriminating variable yields increased power with n=2; but eventually addition of poorly discriminating variables "dilutes" the discrimination power -- e.g., n=6 compared to the best set of 5. (d) Increased dimensionality gives enormous leverage to modest improvements in precision, through the product $\Pi\beta_i$. (See bottom line, Table IV.) These four conclusions directly indicate the way toward improved discrimination power, the last being the most influential. (χ'^2 in the table denotes the non-central χ^2.)

3.3.6 generalization. The foregoing considerations of hypothesis testing and pattern identification limits were necessarily simplified, an extended discussion being beyond the scope of

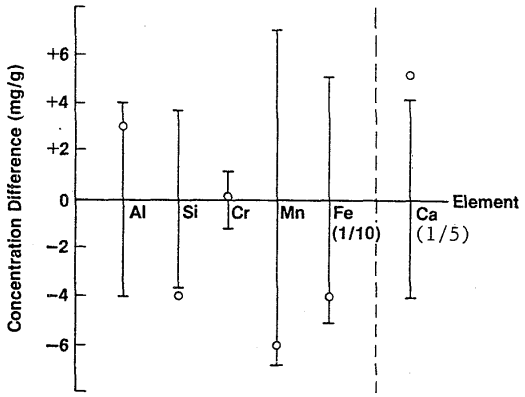

Fig. 13. Multivariable Identification. H_o windows [I] and H_A concentration differences [open circles] for the multi-variable (element) patterns characteristic of particle emissions from two steel plants. [See Table III.]

Table IV. Pattern Discrimination Power [1 - β']

(Steel-A particles vs Steel-B particles [control]; $\alpha' = 0.05$)

	n=1	n=5		n=6
	(Ca)	(-Ca)	(-Cr)	
Sequential matching[a]	0.86	0.87*	0.958	0.949
χ^2 - test[b]	0.88	0.967	0.997	0.996

(a) criterion: $x_i \leq I_i$, all i power: $1 - \Pi \beta_i$
(b) criterion: $\chi^2 \leq \chi^2_c$ power: $1 - P(\chi'^2)$

*8% incr. precision (σ_i's) increases the power to 0.95 [target].

this chapter. The principal generalizations that should be considered, however, are the following:

(1) For the first ("matching") strategy, the requirement of homogeneous variance may be relaxed with the use of individual σ's: i.e., $\sigma_B\sqrt{2}$ for the evaluation of α and I, and $\sqrt{(\sigma_A^2 + \sigma_B^2)}$, to recalculate the β_i (See Figure 12).

(2) For variances estimated as s^2's, t and F would replace z and χ^2, respectively, for hypothesis testing. To estimate the power of the tests, the corresponding non-central distributions would be employed. The non-centrality parameter for the F distribution is the same as for χ^2. This means that even in the best of circumstances (orthogonal variables) this approach to the identification limit or power requires homogeneity of variance and knowledge of σ. (See reference (74) for a discussion of these issues, as well as an in-depth treatment of multivariate hypothesis testing and classification.)

(3) If the H_A universe contains more than one member, its membership and composition must be known for identification to be meaningful. Such knowledge, of course, is in the domain of disciplinary ("scientific") expertise. Proof of H_o [identification] comes only when discriminating power is adequate with respect to all H_A's. For a given control pattern B, only that region of variable space within the discriminating volume need be explored, however. For sequential matching, this means only A - patterns for which the distribution of the difference spectrum A - B significantly overlaps the I-hypercube; for the alternative approach, the discriminating volume derives from the critical value for χ^2. Multiclass discrimination may be performed, for example, through a series of binary tests (12, 17, 74).

(4) Impurity detection for the multivariable case may be treated as a direct extension of the single variable case. For two patterns, the impurity detection limit (component A contaminating control component B) can be calculated from $(\Delta/\sigma_o)_{rms}$ corresponding to $\beta' = 0.05$, where χ^2 ($\alpha' = 0.05$) is used to test the null hypothesis [H_o: $\mathbf{x}_B - \hat{\mathbf{x}}_B$]. For mixed impurities, a "worst case"

limit may be derived from the "pseudopattern" of closest approach. That is, the two pattern discrimination limit is recalculated substituting A′ for A, where pseudopattern A′ is the linear combination of alternative vectors which lies closest to B. At trace levels, observed patterns become increasingly fuzzy, because of measurement imprecision or baseline noise. Clearly, under such circumstances "detection" and "identification" become entwined. (See reference (75).)

(5) Covariance among variables (e.g., elemental concentrations) **within** individual classes is by far the greatest complication. It may be treated in one of four ways: (a) Select just one variable or function of variables (e.g., first principal component); covariance is then undefined. (b) Select only the most powerful, uncorrelated (or nearly independent) variables, discarding others showing significant correlation (12, Chapt 20). (c) Transform the original variables into a reduced, orthogonal set, as in Principal Component Factor Analysis and SIMCA (76,97). (d) In the absence of a very large sample for testing the multivariate normal assumption and estimating the within class covariance matrices, the fourth alternative is daunting: taking into account the full covariance structure through critical contours $[\alpha,\beta]$ of the hyperellipsoids corresponding to H_0 and H_A. Considering just two variables, the treatment would be analogous to the confidence ellipse for the estimated slope and intercept of a fitted calibration line. (Hypothesis testing of calibration curve parameters is far more amenable to this multivariate "parametric" approach, however, since the correlation matrix is known from the design of the experiment (77).) For two variables, the matching intervals I and the respective probabilities $(1-\alpha')$ would not be greatly affected by the lack of rigorous knowledge of the covariance matrix, since $(1-\alpha') \approx (1-\alpha)^2 \approx 1$. The false match probability β' could be significantly in error, however, because $\beta_1 \cdot \beta_2$ must now be replaced by $\beta_1 \cdot (\beta_2|1)$, where $(\beta_2|1)$ is the conditional probability of a false match for variable-2 given a false match for variable-1. If the variables are perfectly correlated, $(\beta_2|1) = 1$, and the second variable lends no incremental discriminating power. Higher dimensions lead to increasing complexity, and estimates of higher order correlations become increasingly imprecise as one runs out of degrees of freedom.

4. CONCLUSIONS AND OUTLOOK

The ability to detect specified (absolute) levels of chemical species in environmental, biological, and physical (material) systems is crucial for the well-being and advancement of our society. Because of the practical importance of reliable detection in the societal setting, on the one hand, and its technical complexity, on the other, we face a "Two Cultures" type situation. We scientists lack the expertise to fully comprehend or effectively influence the sociopolitical issues; experts in that domain, similarly cannot be expected to fully comprehend the technical issues involved. Effective communication and mutual education -- one of the aims of this volume and this overview chapter -- is therefore essential. With this objective in mind, let us re-consider briefly some of the observations and suggestions of this tripartite overview.

4.1 Sociopolitical Perspectives

- Adequate detection capabilities are important to society, both for natural or anthropogenic hazards and for requisite beneficial levels of chemical species -- e.g., nutrients. "Adequacy" means "certainty" to the layman; if a substance is present (above the specified level of concern) it will surely be detected -- the alarm will go off; if not, there will be no (false) alarm.

- Despite the intrinsic uncertainty (false positives and false negatives) associated with detection, and in fact, with all of measurement, the general public is not schooled to accept such a limitation. Ignorance and suspicion with respect to this issue is reflected also when it comes to our ability to reduce concentrations of "bad" species to zero, or for that matter to detect all concentrations exceeding zero (78). Scientific naivete' regarding newly detected noxious species when detection capabilities improve constitutes another form of ignorance having potentially great political impact.

- Sociopolitical "debates," in both the legislative and judicial arenas, have very different ground rules than scientific debates (3). Advocacy, conflicting societal concerns and perceptions, and even "hidden agenda" drive such debates. They cannot, and probably should not, be conducted like a scientific forum. With patience and honest input from the scientific community in its area of expertise, generally the collective common good is served (79,80). With reference to risk management for "dread risks" affecting large numbers of people, for example, Lave observed that collective decisions are mandatory, but because of the diversity of safety goals, collective decisions are difficult (80).

- Risk perceptions and collective (or delegated) decisions lie behind many of our regulatory limits or hazard "alarm levels," [L_R] which, in turn, drive our measurement Detection Limits. Though certain approaches to decision analysis, especially those incorporating Bayesian strategies, might seem appropriate for simultaneously embracing societal risk and measurement error risk (false positives and negatives), it would seem advisable in practice to decouple the two. Let society (or medicine, or affected industry, etc.) enter the political debate to establish their requisite L_R's. Then, Measurement Science, using the appropriate scientific criteria and standards, should attempt to meet these L_R's with scientifically defensible Detection Limits. The late Philip Handler put it well, by stating that "Scientists best serve public policy by living within the ethics of science, not those of politics" (81).

- Societal and scientific perceptions of risk sometimes diverge. Slovic's investigation of ordered risk preferences of laymen vs experts is an interesting case in point (79). Nuclear power, for example, was rated first among representatives of the lay public (League of Women Voters; college students), yet it was 20th in the eyes of experts. Surgery was 5th according to the relevant experts, but it was only 10th in the public view. The nature of our society naturally accords primary weight to that

society's public perceptions, when it comes to political decisions. This deserves our respect for many reasons, including the fact that society's judgment is not constrained by a possibly too narrow view or faulty algorithm. In fact, its "basic conceptualization of risk is much richer than that of the experts and reflects legitimate concerns that are typically omitted from expert risk assessments" (79, p. 285).

- Adequacy of detection limits is something that society has a right to demand, and support if the cost should be high. Inadequate detection capacity for specified levels of fires, earthquakes, toxic organisms, etc. must be addressed through refined sampling and measurement procedures. Inadequate performance of a Measurement Process not only fails to provide sufficient warning, but it may also produce quite misleading conclusions. Elevated levels of Ni in human serum due to occupational exposure (ca 5 ng/mL), for example, were quite undetectable until an excellent reference analytical method was developed under the auspices of the International Union of Pure and Applied Chemistry [IUPAC]. Prior methods, quite incorrectly implied that normal levels of Ni in blood serum were some ten times higher than that occupational exposure level (82).
- The costs of erroneous detection decisions can be quite significant. Disastrous results may follow if irreversible actions are taken. Even the seemingly harmless false positive which can later be shown to be spurious can damage reputations and/or lead to expensive court suits. It is important therefore that scientific detection decisions and detection limits be approached in a quantitative manner, with due attention to the probability of errors of both kinds.

4.2 Technical Issues

- Meaningful detection decisions and detection limits can follow only from rigorous attention to the Measurement Process and an Hypothesis Testing framework for defining detection capability. This is especially appropriate, as hypothesis testing is the expression of the Scientific Method. Decision criteria, detection limits, and acceptable false positive and false negative risks must be quantified, and CMP's designed to meet their specifications. The scientific expertise required goes deep. This was observed, for example, in the investigation of detection limits for a variety of analytical methods for the International Atomic Energy Agency. As illustrated in Ref. 35, detailed, method specific expertise was essential in order to expose certain subtle, but extremely important factors affecting calibration and the blank [Note 7].
- All is not well in the technical camp. Confusion among scientists between the design of the MP to meet requisite levels of performance [L_D], and an experimental outcome or detection decision based on a specified criterion [L_C], is at the heart of much our internal disarray. That is, two different (albeit related) issues are under discussion, often unknowingly and with conflicting terminology. Ad hoc rules of thumb, or simplistic consensus ("voted") formulae are proffered -- often in the

interest of producing a simple ranking of CMP's according to something labeled as an LOD. This serves no one. In particular, it fails to provide the public with meaningful detection capabilities comprising reliable and adequate false positive and false negative error probabilities. Perhaps the most common extreme is the case where the β-error is unrecognized, such that its de facto value is 50% [Note 8].

- The drive toward facile expressions for limits of detection is partly a matter of attitude and education. Solid training in statistics and drilling with respect to the fundamental concepts of experimental design and hypothesis testing in science is missing from the undergraduate education of many chemists in the U.S. Western Europe fares better; and now that training in Chemometrics is beginning to appear in the American curricula (83), real hope exists for common understanding of these matters by the "ordinary" chemists of the future. An illustration of the present state comes from a survey recently taken by an instrument manufacturer of its users in the nuclear industry. Regarding topical material covered at workshops, comments came back that users would prefer omission of the theory with more time spent on use of the formulas. A personal view is that education related to basic concepts should always have the priority; understanding (and questioning of) formulas is important, but calculators or computers are quite proficient at using them.

- The link between "ordinary" measurements and detection limits needs reinforcing. That is, both depend for their validity on all sources of systematic and random error associated with the entire CMP. Thus, for example, detection decisions [tests of significance] and confidence intervals depend on the same assumptions and error components for their validity. If Student's t is appropriate to use with the one, it is equally appropriate for the other.

- Conventions for reporting data, and "black box" algorithms can induce subtle bias into many types of modern chemical/instrumental data, but the problem is exacerbated with the growth of automatic laboratory systems and low level measurements and data bases. The black box may contain mistakes, and all too often its mechanism is unavailable to the user, and on occasion that mechanism (i.e., algorithm) changes for low level observations. Information loss or distortion, whether it occurs within the black box or by the pen of the user, is especially severe for low level data. Its impact on long term storage and data base generation is an issue of some importance (34).

- Quality control at low levels (blank, detection limit) must be addressed both with Standard Reference Materials and Standard Test Data, if we are to certify the accuracy of our detection decisions and detection limits. Since the blank has such a profound influence on the validity of detection decisions, it deserves special attention. The CMP must be designed to incorporate an adequate number of "real" blanks, and it should take advantage of the normalizing tendency of averages from paired comparisons.

- The introduction of Discrimination Limits, such that small non-zero concentrations will rarely produce false positives, should do much to alleviate the public alarm that sometimes follows such "detection." At the same time it could avert the common implicit overcompensation associated with ignoring of the error of the second kind [false negative]. Also, those who decry current usage of detection limits because they are too imprecise, or equivalent to the flipping of a coin, might regard Discrimination Limits as useful, more precise measures of detection capability, still in keeping with the hypothesis testing concept.
- Discrimination Limits and multiple detection decisions lead naturally to univariate and multivariate formulations for identification, an outgrowth of the fundamental concept of hypothesis testing. Methods for treating this link have been developed, so it becomes natural at this point for us to address **together the two primary characteristics of Qualitative Analysis**: Detection and Identification.
- Identification differs in one, very critical respect from detection: a consistency test of the null hypothesis is necessary but **not** sufficient for identification. Discrimination limits must be adequate for **all** alternative hypotheses (other substances). At this point scientific intuition or expertise plays a crucial role, for we must somehow discover the universe of all possible alternatives to the substance we wish to identify, in the context of the given measurement process.

4.3 Pre- and Post-History: The Challenge. The concept of the Detection Limit, at least in Analytical Chemistry, was slow to evolve in the early decades of this century from a loose, qualitative idea, to a potentially semi-rigorous numerical attribute for a fully-defined CMP. During the past twenty years or so, important strides have been made in education and in the development of a consistent and practically useful formulation of the Detection Limit, especially in Europe. Unfortunately, diversity in understanding, formulation, and nomenclature among scientists continues. This has been exacerbated by the demand for regulations and simplified rules and formulas, often on relatively short time scales. "Definitions" deriving from polemics or from democratic, consensus tactics are unlikely to meet long term standards for scientific rigor (conceptual rigor, not necessarily uncertainty-free, numerical rigidity).

Although a sound approach to detection has been available for at least two decades, and despite its current successful application to many practical and scientific problems, the current disarray among scientists in the U.S. [cf Fig. 4] can only further mystify the public in an area that seems already inherently mystical. The promise comes from trends in chemical education and from work in progress in reputable international chemical organizations. Statistics and the proper concepts of measurement uncertainty, experimental design, and hypothesis testing are gaining a foothold in the undergraduate chemical curriculum, especially under the stimuli of modern instrumental and computational facilities and Chemometrics (84). Also, at the present time at least two

commissions of IUPAC, partly in collaboration with the international chemometrics community, are drafting guidelines and nomenclature documents treating a broad range of chemical measurement issues, including those related to uncertainty, experiment design, reporting of data, and detection.

Cooperation between the two cultures should become increasingly fruitful, as common concern in meeting society's legimitate needs for practical detection capabilities bind us together, and as we each invest our efforts in our respective areas of expertise. Mutual education and inter-cultural communication can only accelerate this process.

Acknowledgment

Special thanks go to the following colleagues, for their important suggestions and care in reading a draft of this chapter: K. R. Eberhardt, M. S. Epstein, R. W. Gerlach, H. M. Kingston, P. A. Pella, C. H. Spiegelman, R. A. Velapoldi, and J. W. Winchester.

Literature Cited

1. Kutschera, W. "Rare Particles"; Nucl. Instrum. Meth. B5, 1984, 233, 420.
2. Cooper, R. M. "Stretching Delaney Till It Breaks"; Regulation, Nov/Dec 1985, 11.
3. Moss, T. "Scientific Measurements and Data in Public Policy Making"; Chapt. 3 in this volume.
4. Science, Risk Assessment Issue, 1987, 236, 267-300.
5. Currie, L.; Klouda, G.; Voorhees, K. "Atmospheric Carbon"; Nucl. Instrum. Meth., B5, 1984, 233, 371.
6. Rogers, L. B. "Interlaboratory Aspects of Detection Limits Used for Regulatory/Control Purposes"; Chapt. 5 in this volume.
7. Rensberger, B. "A Life is Worth $2 Million, Regulatory Analysis Shows"; Science Notebook, Wash. Post, Mar. 2, 1987. [Science news summarizing highlights of a study to be published in Environmental Science and Technology.]
8. Liteanu, C.; Rica, I. Statistical Theory and Methodology of Trace Analysis. New York: John Wiley & Sons; 1980.
9. Egan, J. P. Signal Detection Theory and ROC Analysis, Academic Press, New York, 1975.
10. Frank, I. E.; Pungor, E.; Veress, G. E. "Statistical Decision Theory Applied to Analytical Chemistry"; Anal. Chim. Acta 133 (1981) 433.
11. Howard, R. A. "Decision Analysis: Perspectives on Inference, Decision, and Experimentation"; Proc. IEEE, 1970, 58, 823.
12. Massart, D. L.; Dijkstra, A.; Kaufman, L. Evaluation and Optimization of Laboratory Methods and Analytical Procedures, New York, Elsevier, 1978.
13. Currie, L. A. The Discovery of Errors in the Detection of Trace Components in Gamma Spectral Analysis, in Modern Trends in Activation Analysis, Vol. II. J. R. DeVoe; P. D. LaFleur, Eds.; Nat. Bur. Stand. (U.S.) Spec. Publ. 312; p. 1215, 1968.

14. Kirchmer, C. J. "The Estimation of Limit of Detection for Environmental Analytical Procedures" - Chapt. 4 in this volume.
15. Boumans, P. W. J. M. A Tutorial Review of Some Elementary Concepts in the Statistical Evaluation of Trace Element Measurements. Spectrochim. Acta 33B: 625; 1978.
16. Keith, L. H.; Crummett, W; Deegan Jr, J; Libby, R. A.; Taylor, J. K.; Wentler, G. "Principles of Environmental Analysis", Analyt. Chem. 1983, 55, 2210.
17. Kateman, G.; Pijpers, F. W. Quality Control in Analytical Chemistry. New York: John Wiley & Sons; 1981.
18. Heydorn, K.; Wanscher, B. Application of Statistical Methods to Activation Analytical Results Near the Limit of Detection. Fresenius' Z. Anal. Chem. 292(1): 34-38; 1978, See also: Heydorn, K. Neutron Activation Analysis for Clinical Trace Element Research, 2 Vol., Boca Raton: CRC Press; 1984.
19. Feigl, F. Tüpfel- und Farbreaktionen als mikrochemische Arbeits-methoden, Mikrochemie 1: 4-11; 1923.
20. Kaiser, H. Z. Anal. Chem. 209: 1; 1965 [Ref. 32].
Kaiser, H. Z. Anal. Chem. 216: 80; 1966.
Kaiser, H. Two Papers on the Limit of Detection of a Complete Analytical Procedure, English translation of the above manuscripts. London: Hilger; 1968.
21. Altshuler, B.; Pasternack, B. Statistical Measures of the Lower Limit of Detection of a Radioactivity Counter. Health Physics 9: 293-298; 1963.
22. Wilson, A. L. The Performance-Characteristics of Analytical Methods. Talanta 17: 21; 1970; 17; 31: 1970; 20: 725; 1973; and 21: 1109; 1974.
23. Currie, L. A. Limits for Qualitative Detection and Quantitative Determination. Anal. Chem. 40(3): 586; 1968.
24. St. John, P. A.; Winefordner, J. D. A Statistical Method for Evaluation of Limiting Detectable Sample Concentrations. Anal. Chem. 39: 1495-1497; 1967.
25. Nicholson, W. L. "What Can Be Detected"; Developments in Applied Spectroscopy, v.6, Plenum Press, p. 101-113, 1968; Nicholson, W. L., Nucleonics, 24 (1966) 118.
26. Grinzaid, E. L.; Zil'bershtein, Kh. I.; Nadezhina, L. S.; Yufa, B. Ya. Terms and Methods of Estimating Detection Limits in Various Analytical Methods. J. Anal. Chem. - USSR 32: 1678; 1977.
27. Ingle, J. D., Jr. Sensitivity and Limit of Detection in Quantitative Spectrometric Methods. J. Chem. Educ. 51(2): 100-5; 1974.
28. Currie, L. A. "Lower Limit of Detection: Definition and Elaboration of a Proposed Position for Radiological Effluent and Environmental Measurements," U S Nuclear Regulatory Commission, NUREG/CR-4007, 1984.
29. IUPAC Comm. on Spectrochem. and other Optical Procedures for Analysis. Nomenclature, symbols, units, and their usage in spectrochemical analysis, Pure Appl. Chem. 45 (1976) 99.
30. IUPAC, Commission V.3, Recommendations for Nomenclature in Evaluation of Analytical Methods, Draft Report, 1986.
31. Kaiser, H. Spectrochim. Acta, 1947, 3, 40.
32. Kaiser, H. Z. Anal. Chem., 1965, 209, 1.
33. W. B. Crummett in Ref. 71.

34. Brossman, M. W., Kahn, H; King; D.; Kleopfer, R.; McKenna; G.; Taylor, J. K. Reporting of Low-Level Data for Computerized Data Bases - Chapt. 17 in this volume.
35. Currie, L. A.; Parr, R. M. "Perspectives on Detection Limits for Nuclear Measurements in Selected National (US) and International (IAEA) Programs" - Chapt. 9 in this volume.
36. Cheeseman, R. V.; Wilson, A. L. Manual on Analytical Quality-Control for the Water Industry - Relating to the Concept of Limit of Detection, WRc Environment, Water Research Center, Medmenham, UK, 1978.
37. Ramos, L. S.; Beebe, K. R.; Carey, W. P.; Sanchez, M. E.; Erickson, B. C.; Wilson, B. E.; Wangen, L. E.; Kowalski, B. R. "Chemometrics"; Anal. Chem., 1986, 58, 294R. [Review and bibliography].
38. Long, G. L.; Winefordner, J. D. "Limit of Detection: A closer look at the IUPAC definition"; Anal. Chem., 1983; 55; 712A.
39. Natrella, M. G. 'The Relation between Confidence Intervals and Tests of Significance'; in Ku, H, Ed. "NBS Spec Publ 300"; 1969.
40. Smit, H. C.; Steigstra, H. "Noise and Detection Limits in Signal Integrating Analytical Methods"; - Chapt. 7 in this volume.
41. Epstein, M. S. "Comparison of Detection Limits in Atomic Spectroscopic Methods of Analysis"; - Chapt. 6 in this volume.
42. Ku, H. H. Edit., Precision Measurement and Calibration, NBS Spec. Public. 300 (1969), p. 315.
43. Saw, J. G.; Yang, M. C. K.; Mo, T. C. 'Chebyshev Inequality with Estimated Mean and Variance'; The Amer. Statistician; 1984; 38; 130.
44. Johnson, J. E.; Johnson J. A. "Radioactivity Analyses and Detection Limit Problems of Environmental Surveillance at a Gas-Cooled Reactor" - Chapt. 14 in this volume.
45. Snedecor, G. W.; Cochran, W. G. Statistical Methods, 6th Edit., Iowa State Univ. Press. (1973).
46. Kingston, H. M.; Greenberg, R. R.; Beary, E. S.; Hardas, B. R.; Moody, J. R.; Rains, T. C.; Liggett, W. S. "The Characterization of the Chesapeake Bay: A Systematic Analysis of Toxic Trace Elements"; National Bureau of Standards, Washington, DC, 1983, NBSIR 83-2698.
47. Scales, B. Anal. Biochem. 5, 489-496 (1963).
48. Currie, L. A. "Model Uncertainty and Bias in the Evaluation of Nuclear Spectra"; J. of Radioanalytical Chemistry 39, 223-237 (1977).
49. Koch, W.; Liggett, W. "Critical Assessment of Detection Limits for Ion Chromatography" - Chapt. 11 in this volume.
50. Watters, R. L.; Wood, L. J. in Ref. 71.
51. Murphy, T. J. The Role of the Analytical Blank in Accurate Trace Analysis, NBS Spec. Publ. 422, Vol. II, U. S. Government Printing Office, Washington, DC, 509 (1976).
52. Kelly, W. R.; Hotes, S. A. "The Importance of Chemical Blanks and Chemical Yields in Accurate Chemical Analysis"; Preprint (1987).

53. Kelly, W. R.; Fassett, J. D.; Hotes, S. A., 'Determining Picogram Quantities of U in Human Urine by Thermal Ionization Mass Spectrometry'; Health Physics, 1987; 52, 331.
54. Currie, L. A. 'Quality of Analytical Results, with Special Reference to Trace Analysis and Sociochemical Problems'; Pure & Appl. Chem., 1982, 54, 715.
55. Currie, L. A. The Limit of Precision in Nuclear and Analytical Chemistry. Nucl. Instr. Meth. 100: 387; 1972.
56. Lub, T. T.; Smit, H. C., Anal. Chim. Acta, 1979, 112, 341.
57. Mellor, R. A.; Harrington, C. L. "Evaluating the Impact of Hypothesis Testing on Radioactivity Measurement Programs at a Nuclear Power Facility" - Chapt. 13 in this volume.
58. Parr, R. M.; Houtermans, H.; Schaerf, K. 'The IAEA Intercomparison of Methods for Processing Ge(Li) Gamma-Ray Spectra'; in "Computers in Activation Analysis and Gamma-Ray Spectroscopy"; U.S. Dept. of Energy, Sympos. Ser. 49, 1979, 544.
59. Liggett, W. ASTM Conf. on Quality Assurance for Environmental Measurements, 1984, Boulder, CO.
60. Watters, R. L.; Spiegelman, C. H., and Carroll, R. J. "Heteroscedastic Calibration in Inductively Coupled Plasma Spectrometry"; Anal. Chem., 1987, 59, 1639.
61. Currie, L. A. "The Many Dimensions of Detection in Chemical Analysis"; in Chemometrics in Pesticide/Environmental Residue Analytical Determinations, ACS Sympos. Series (1984).
62. Owens, K. G.; Bauer, C. F.; Grant, C. L. "Effects of Analytical Calibration Models on Detection Limit Estimates";- Chapt. 10 in this volume.
63. Iyengar, G. V.; Tanner, J. T.; Wolf, W. R.; Zeisler, R. 'Preparation of a Mixed Human Diet Material for the Determination of Nutrient Elements, Selected Toxic Elements and Organic Nutrients: a Preliminary Report', submitted to The Science of the Total Environment, 1986.
64. Currie, L. A. Detection and Quantitation in X-ray Fluorescence Spectrometry, Chapter 25, X-ray Fluorescence Analysis of Environmental Samples, T. Dzubay, Ed., Ann Arbor Science Publishers, Inc., p. 289-305 (1977).
65. Horwitz, W.; in Ref. 71.
66. Zweig, M. "Establishing Clinical Detection Limits of Laboratory Tests"; - Chapt. 8 in this volume.
67. Natrella, M. G.; Experimental Statistics, NBS Handbook 91, 1963.
68. Dixon, W. J.; Massey, F. J. Introduction to Statistical Analysis, McGraw-Hill, New York, 1957.
69. Schlain, B., personal communication, 1987.
70. Currie, L. A. 'The Limitations of Models and Measurements as Revealed through Chemometric Intercomparison'; J. Res. NBS, 1985, 90, 409.
71. Kurtz, D.; Taylor, J. K.; Sturdivan, L.; Crummett, W.; Midkiff, C.; Watters, R.; Wood, L.; Hanneman, W.; Horwitz, W. Real-World Limitations to Detection: A Panel Discussion; Chapt. 16 in this volume.
72. Currie, L. A.; Gerlach, R. W.; Lewis, C. W.; Balfour, W. D.; Cooper, J. A.; Dattner, S. L.; DeCesar, R. T.; Gordon, G. E.; Heisler, S. L.; Hopke, R. K.; Shah, J. J.; Thurston, G. D.;

Williamson, H. J. "Interlaboratory Comparison of Source Apportionment Procedures: Results for Simulated Data Sets"; Atmospheric Environment 18 (1984), 1517.
73. Eisenhart, C.; Zelen, M. Ch 12, "Elements of Probability"; in Handbook of Physics, E. U. Condon and H. Odishaw, Ed., McGraw-Hill, New York, 1958.
74. Kendall, M.; Stuart, A.; Ord, J. K. The Advanced Theory of Statistics; Vol. 3, MacMillan; New York,; 1983.
75. Delaney, M. F. 'Multivariate Detection Limits for Selected Ion Monitoring Gas Chromatography - Mass Spectrometry'; Chemometrics and Intelligent Laboratory Systems, 1987.
76. Wold, S.; Albano, C.; Dunn, III, W. J.; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjöström, M. "Multivariate Data Analysis in Chemistry"; in Chemometrics: Mathematics and Statistics; B. R. Kowalski, Ed., (Reidel Publishing Co.) 1984; pp. 17-96.
77. Draper, N.; Smith, H. Applied Regression Analysis, Wiley New York, 1981.
78. McCormack, M. "Realistic Detection Limits and the Political World" - Chapt. 2 in this volume.
79. Slovic, P. "Perception of Risk"; 280-285 in Ref. 4.
80. Lave, L. "Health and Safety Risk Analysis: Information for Better Decisions"; 291-295 in Ref. 4.
81. Handler, P. Dedication Address, Northwestern Univ. Cancer Center, 1979. (See also S. J. Gould in the New York Times Magazine, 19 April 1987.)
82. Nieboer, E.; Jusys, A. A. "Contamination Control in Routine Ultratrace Analysis of Toxic Metals"; Ch. 1 in Chemical Toxicology and Clinical Chemistry of Metals, S. S. Brown and J. Savory, Eds., Academic Press, London, 1983.
83. Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. Chemometrics, Wiley, New York, 1986.
84. Kowalski, B. R., Ed. Chemometrics: Mathematics and Statistics in Chemistry, (Reidel Publishing Co.) 1984.
85. Hurst, G. S.; Payne, M. G.; Kramer, S. D.; Young, J. P. Rev. Mod. Phys., 1979, 51, 767.
86. Donaldson, W. T., Envir. Sci. Tech., 1977, 11, 348.
87. Guinn, V., personal communication; 1987.
88. Donahue, D., Fourth Intern. Sympos. on Accelerator Mass Spectrometry, Niagara on the Lake, 1987.
89. Heydorn, K.; Christensen, L. H. "Verification Testing of Commercially Available Computer Programs for Photpeak Area Evaluation"; Intern. Conf. on Meth. Appl. Radioanalytical Chemistry, Kona, 1987.
90. Cochran, W. G.; Biometrics, 1964, 20, 191.
91. Andrews, R. M. "Meat Inspector: 'Eat at Own Risk'"; Wash. Post, May 16, 1987. [Science news following issuance of a National Academy of Sciences Report concerning food poisoning from undetected microorganisms.]
92. Lochamy, J. D. The minimum-detectable-activity concept. Nat. Bur. Stand. (U.S.) Spec. Publ. 456; 1976, 169-172.
93. IAEA; Users' Guide on Limit of Detection; in preparation. (See Ref. 35).

94. "Standard Radiological Effluent Technical Specifications for Pressurized Water Reactors," U. S. Nuclear Regulatory Commission, NUREG-0472, Rev. 3, September 1982.
95. Federal Register, 1984, 49, 43431.
96. Federal Register, 1984, 49, 43430; and Glaser, J., Foerst, D., McKee, G., Quave, S., and Budde, W., "Trace Analysis for Wastewaters," Environ. Sci. Tech., 1981, 15, 1426.
97. Forina, M.; Lanteri, S. "Data Analysis in Food Chemistry;" pp. 305-349; in Ref. 84.
98. Feigl, F. Chemistry of Specific, Selective and Sensitive Reactions; New York: Academic Press; 1949.
99. Emich, F. Ber. 1910; 43; 10.
100. Cox, D. R.; Lewis, P. A. W. The Statistical Analysis of Series of Events; London: Methuen; 1966.

Notes

Note 1. Analytical advances have led to the possibility of "single atom detection" (85). At the same time it is recognized that at concentrations of 1 part in 10^{15} (in water) in principle "every known organic compound could be detected" (86). These measurement realities mandate the setting of regulatory levels on bases other than either non-zero concentrations, or the inherent ability to detect.

Note 2. That Feigl's "Identification Limit" referred to the minimum quantity detectable (L_D) as opposed to the decision or critical level (L_C) is clear from his statement defining the *"'Erfassungsgrenze'* [as] *die Kleinste absolute Menge Substanz ... die ... noch nachweisbar und bestimmbar ist* "(Ref. 19, p. 6). In a later, english language publication, this meaning was amplified in a manner that foreshadowed the modern statistical approach to detection. In the volume "Chemistry of Specific, Selective, and Sensitive Reactions", p. 14 (98), Feigl described a test for magnesium which was "always" positive, for 40 repetitions, using a 0.05% Mg solution. With dilution by factors of 10 and 50, however, the test was positive only in 24 and 6 instances, respectively. With this, Feigl embraced the concept of the "region of uncertain reaction" (99), and a condition for the identification limit that the chance of a false negative be negligible.

Note 3. Symbols introduced in this section include the following: y [gross signal], B [null signal = background, baseline, or blank], S [net signal], x [analyte concentration or amount], A ["sensitivity" or calibration factor], pdf [probability density function], cdf [cumulative distribution function], superscript ^ or est( ) [estimated value], E() or μ [expected value], V or σ^2 [population variance], s^2 [estimated variance], σ_o [standard deviation of the estimated net signal, when E(S)=0], CI [confidence interval], df [degrees of freedom], Δ [bias], Δ_D [bias detection limit; discrimination limit]. Subscripts, $_m$, $_M$, denote lower and upper limits, respectively.